THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXIX MAY 1960 NUMBER 3

Capabilities of the Telephone Network for Data Transmission
A. A. ALEXANDER, R. M. GRYB AND D. W. NAST 431

High-Frequency Negative-Resistance Circuit Principles for Esaki Diode Applications
M. E. HINES 477

Theory of Current-Carrier Transport and Photoconductivity in Semiconductors with Trapping
W. VAN ROOSBROECK 515

The Charge and Potential Distributions at the Zinc Oxide Electrode
J. F. DEWALD 615

The Square of a Tree
I. C. ROSS AND F. HARARY 641

Theory of a Frequency-Synthesizing Network
B. M. WOJCIECHOWSKI 649

An Evaluation of AM Data System Performance by Computer Simulation
R. A. GIBBY 675

Semiconductor Strain Transducers
F. T. GEYLING AND J. J. FORST 705

Recent Bell System Monographs 733

Contributors to This Issue 742

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. F. R. Kappel, President; S. Whitney Landon, Secretary; L. Chester May, Treasurer. Subscriptions are accepted at $5.00 per year. Single copies $1.25 each. Foreign postage is $1.08 per year or 18 cents per copy. Printed in U.S.A.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXIX MAY 1960 NUMBER 3

Capabilities of the Telephone Network for Data Transmission

By A. A. ALEXANDER, R. M. GRYB and D. W. NAST

(Manuscript received January 22, 1960)

This paper presents the results of a nationwide data transmission field testing program on the telephone switched message network. Error performance using the FM digital subset is described and basic transmission characteristics such as net loss, bandwidth, envelope delay and noise are given.

I. INTRODUCTION

The telephone industry has a long history of providing data transmission services. Over the years many varieties of service offerings making use of the range from narrow-band telegraph[1] channels up to 4-mc video channels have been provided. Such services usually have been provided on a private line basis by adapting regular telephone facilities to the particular data service requirement.

The increased use of computers and automatic data processing systems in the commercial, industrial and military areas has substantially increased the demand for greater varieties of data services and data transmission channels. This expansion, with its attendant requirement for a variety of speeds and channel usage time, has encouraged development of service offerings that use the regular switched message telephone network in establishing the communication channels. In the Bell System this service concept has been given the name of Data-Phone,* which

* Data-Phone is a trademark of the American Telephone and Telegraph Company identifying Bell System equipment used in this Bell System service.

envisions not one, but a whole family of data transmission systems operating on the regular switched network. It will encompass a broad range of speed capabilities and meet a variety of performance requirements.

Operationally, the Data-Phone service is quite simple. A regular telephone call is made to establish a connection between two points. Usually, regular voice communication may be carried on if required. Operation of a pushbutton associated with the telephone set at each end of the connection disconnects the telephone instruments and connects data subsets to the telephone lines. The subset, depending upon the type, accepts analog or digital (usually binary) information at the transmitting end and, if necessary, modulates the baseband signal to a frequency band suitable for use over telephone circuits. At the receiving end the data subset demodulates the line signal and returns it to baseband. At the end of the transmission regular voice communication can be resumed, if required, or the connection can be terminated by hanging up the telephone set.

Additional operational features can, of course, be built into the Data-Phone service as may be required. For example, machines may be used to dial up the connection, answer back, intercommunicate and disconnect entirely independent of human assistance. These and other similar features are obvious extensions of the Data-Phone concept.

The switched telephone network is designed primarily to handle voice communication. Many of the design criteria are based upon talker and listener habits and preferences. The resulting characteristics, while suitable for data transmission are not as optimized as they are when a communication channel is designed specifically for data use. Frequently it is possible to take advantage of certain speech or human ear characteristics to provide better or more economical service. The uses of compandored[2] carrier systems and echo suppressors[3] are typical examples. When the telephone network is used to provide communication channels for systems having nonhuman characteristics the advantages of these special devices are lost.

In order to use the telephone network for the variety of uses contemplated under the Data-Phone concept, it is necessary to know the following:

i. What is contained in the telephone network and how it operates.

ii. What are the voice and data transmission characteristics of connections in the message network and to what extent they limit the transmission of data signals.

Once these are determined an objective evaluation of the switched message network can be made, and data systems can be designed with a reasonable degree of assurance for successful application.

II. SWITCHED TELEPHONE NETWORK

A great deal of information describing the component parts and operating characteristics of the switched message network has been published. Refs. 2 through 11 describe some of the significant operating and engineering features.

The connections that are established in completing telephone calls show a very large variation in characteristics that are of importance to the transmission of data signals. This stems primarily from two factors:

i. There are a large variety of transmission systems used in the telephone plant (see Refs. 12 through 22). Table I lists some of the more important ones used in the Bell System today.

ii. The number of switched links (trunks) that are used to make up a given connection is quite variable. It is significant that a given long distance call may have as many as nine trunks switched in an over-all connection, or it may have as few as three. Two telephone calls between the same places may go over entirely different routes, pass through different offices and use different numbers of switched trunks.

The system characteristics usually of interest for data transmission include amplitude-frequency response, envelope delay-frequency characteristic, net loss, noise and echo suppressor turnaround time. In the case of voice-frequency cable, loaded or nonloaded, the characteristics are a function of length of loop or trunk. In a carrier system, the noise performance is a function of length and repeater spacing. As a practical matter, attenuation, envelope delay and noise also vary somewhat between channels of the same carrier system.

There are many cases where, for one reason or another, a particular trunk between switching offices is made up of a number of different transmission systems. The resultant trunk characteristics are then a combination of the characteristics of several systems.

In order to relieve the Data-Phone customer of the responsibility for engineering to accommodate the variability in transmission characteristics, subsets (usually modems with various control features) are provided. These act as buffers between customer data-generating or data-using equipment and telephone line characteristics, and provide well-defined interface arrangements.

III. DESCRIPTION OF THE FIELD MEASUREMENT PROGRAM

Data transmission characteristics of telephone connections have a wide range of variability. On the same basis, the error performance of Data-Phone systems might be expected to be quite variable and dependent on a large variety of conditions. An evaluation of data performance in the

TABLE I—BELL SYSTEM TRANSMISSION SYSTEMS USED ON MESSAGE TRUNKS

Type of Transmission System	Transmission Medium	Primary Application	Degree of Use	Remarks
Voice, open wire	Wire	Interlocal office trunks and class 5 to higher class offices	Small	Mostly rural application
Nonloaded voice cable	Cable	Interlocal office trunks and class 5 to higher class offices	Large	In most cases lengths are short (few thousand feet)
Loaded voice cable, all types[13]	Cable	Interlocal office trunks and class 5 to higher class offices	Large	
Type C carrier[14]	Open wire	Between higher class office	Medium	
Type H carrier[15]	Open wire	Class 5 to higher class offices and short haul between higher class offices	Small	
J carrier[16]	Open wire	Between higher class offices	Small	
M carrier	Open wire or power line	Class 5 to higher offices	Small	Also used for telephone loops in rural areas
K carrier[17]	Cable	Between higher class offices	Large	
L carrier[18]	Coaxial cable	Between higher class offices	Large	
N carrier[19]	Cable	Interlocal offices, class 5 to higher class offices and short haul between higher class offices	Large	
O carrier[20]	Open wire	Between higher class offices	Large	
ON carrier[21]	Cable	Interlocal office trunks, class 5 to higher class offices and short haul between higher class offices	Large	
TD system[22]	Radio	Between higher class offices	Large	
TJ system	Radio	Between higher class offices, short haul	Small	Large degree of use expected
TH system	Radio	Between higher class offices		Not yet in service — large degree of use expected

face of such variability leads inevitably to the use of sampling techniques and descriptions in terms of statistical distributions and probabilities. A field testing program designed to sample this variety of conditions and situations has been undertaken.

The objectives of the field testing program on the switched message network were to determine the following:

i. the statistics of error performance, permitting evaluation of error detection and error correction techniques where necessary;

ii. the factors that cause error to occur;

iii. the data speed capabilities and practical operating conditions, and the factors that limit them;

iv. the statistics concerning basic transmission characteristics — this is invaluable in designing new or improved data transmission systems.

Arrangements were made to place calls, send data signals and measure transmission parameters and error performance. Teams equipped with mobile testing terminals made telephone calls between varieties of locations throughout the country. The locations selected were places where potential Data-Phone customers were most likely to be found—in business districts, commercial areas and suburban industrial sites. Testing was carried out within, around and between New York, Chicago, Dallas, San Francisco and Los Angeles. These areas were selected as representative of the variety of conditions and facilities that exist in the present telephone network.

In the program about 1100 test calls were made. About 25 per cent of these were local calls not involved with the long distance switching plan. About 25 per cent were short-haul long distance calls, of up to about 400 miles airline distance. The remaining 50 per cent were long haul, 400 to 3000 miles long.

In order to keep the testing program within manageable size, a single data transmitting system known as the FM digital subset[23] was used for the higher-speed data performance tests. In this system the modulator accepts baseband binary information in serial form and provides a frequency-modulated output. The marking condition is one frequency, the spacing condition another. A single oscillator swings between the two frequencies and transmits the binary information to the demodulator.

The demodulator is a zero-crossing detector, pulse generator and integrator, which provides serial binary baseband signals at its output. The output signals are reproductions of the modulator input signals modified by distortion effects of the telephone facility and modulation-demodulation process.

The error statistics of the telephone network that were determined

TABLE II—TRANSMISSION MEASUREMENTS

Type of Measurement	Conditions of Measurement	Measuring Equipment
Amplitude-frequency response	Between 600-ohm terminations at intervals of about 200 cps	Western Electric Co. 21A transmission measuring set
Envelope delay-frequency response	Between 600-ohm terminations at intervals of about 200 cps.	Acton Laboratories 451 & 452 envelope delay set
Steady noise	F1A weighting	Western Electric Co. 2B noise measuring set
Impulse noise — two methods	Counts in 30 minutes above given power levels, 144 weighting	Western Electric Co. 2B noise measuring set and General Radio 1556-A impact noise analyzer
	Counts in 30 minutes above given power levels, data system band filter weighting	Electronic slicer and counter
Noise recording	Unweighted 5-kc band	Ampex Model 307 magnetic tape recorder

with the FM modem reflect the characteristics of that data system. Measurements taken with some other type of modulation system would probably be somewhat different. In order to minimize the need for testing other types of systems under the conditions encountered during the tests, basic transmission characteristic measurements were also made on

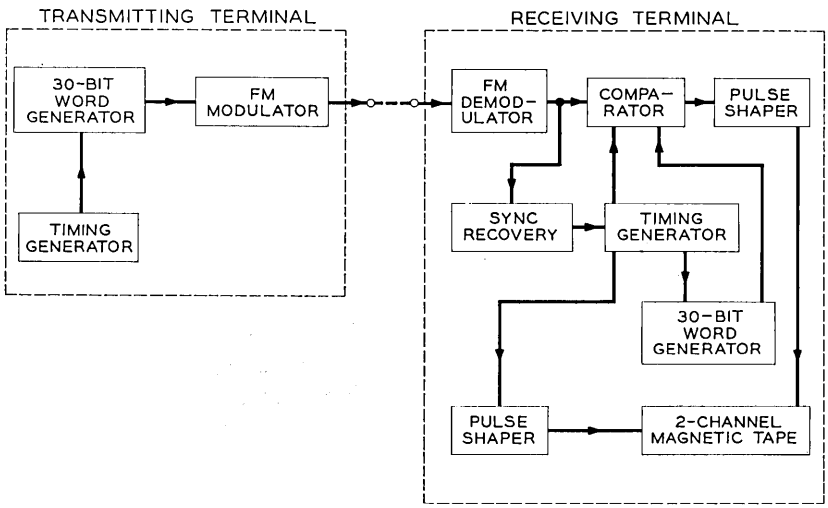


Fig. 1 — Block diagram of transmitting and receiving data terminals.

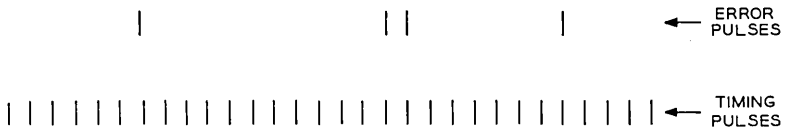


Fig. 2 — Method of error recording on magnetic tape.

each connection. At a later time these conditions may be simulated in the laboratory and comparisons made between the FM system and other modulation systems. A list of the more significant measurements is shown in Table II.

A block diagram showing the transmitting and receiving data terminals with associated circuitry is shown in Fig. 1. A 30-bit word generator, timed by a tuning fork clock,[24] was used to drive the FM subset modulator at the transmitting end. At the receiving end the line signal was demodulated and fed into a comparator. There the binary signal was compared bit by bit with a locally generated 30-bit word. A sync-deriving circuit provided timing information for the receiving terminal clock. The comparator circuit provided an output voltage spike whenever the demodulated 30-bit word and the locally generated 30-bit word did not compare on a bit-by-bit basis. This error indication was recorded on one track of a two-channel magnetic tape; the other channel recorded the locally generated timing signal. This method of recording the time and error information as bits in error and good bits between bits in error (see Fig. 2) permitted later analysis of error statistics by machine methods. The coding of the 30-bit word that was used for the error rate tests is shown in Fig. 3. This coding was selected to provide representative limiting conditions.

Additional circuitry and equipment were provided to permit varying the operating conditions of the data terminals. One of the limiting factors on data performance is the signal-to-noise ratio at the receiving end. Other things being constant, this is directly proportional to the signal level at the transmitting end. The transmitting level should, of course, be as high as is consistent with satisfactory operation on the telephone facility.

The maximum level permissible on telephone facilities is limited by two considerations:

Fig. 3 — Coding of 30-bit word.

i. the coupling loss to other circuits operating in the same cable, open wire or carrier system;

ii. the power-handling capacity of carrier or radio system grouping amplifiers or modulators — overloading may cause modulation products to be generated that will fall in other channels of the same system.

For the tests, a level of −6 dbm at the transmitting terminal was selected to meet established criteria for interference and overloading. Means were provided to reduce the output level in discrete steps, so that relationships between error rate and transmitting level might be determined.

Based upon previous studies and experience, it was determined that the most satisfactory operating region was likely to be centered somewhere between 1200 and 1800 cps and that, at least initially, a speed of 600 bits per second should be used for these tests. Provision was made to operate the data terminals at three pairs of mark-space frequencies: 900(M)–1400(S), 1400(M)–1900(S) and 1100(M)–1900(S).

Information gathered during the early part of the program indicated that sufficient margin was available to permit increasing the speed if the effect of amplitude and delay distortion could be reduced. Compromise amplitude and delay equalizers were designed, and the latter part of the program was carried out using a speed of 1200 bits per second, with mark-space frequency pairs of 1100(M)–2100(S), 1200(M)–2200(S) and 1300(M)–2300(S).

In order to accommodate the increased frequency spectrum, the digital subject was modified with a more optimum bandpass filter and integrating filter.

Error rate information was taken at 600 bits per second using the 900(M)–1400(S) frequency pair. At 1200 bits per second, the 1100(M)–2100(S) pair was used. The three frequency pairs at each speed were used to determine the best operating region for each connection. This was done by measuring maximum repetitive jitter in the transitions of the 30-bit word binary signal as received at the output of the demodulator. (Jitter is the total variation in time of the binary transitions from what they should be; the timing standard is supplied by the receiving clock.) The peak jitter may be expressed in terms of per cent peak distortion in accordance with the following (as shown on Fig. 4):

$$\text{per cent peak distortion} = \frac{\text{max. variation in transition time}}{\text{time of two bits}} \times 100.$$

(The maximum possible distortion is 50 per cent.)

The per cent peak distortion (repetitive jitter) was used as the criteria for determining the best pair of operating frequencies.

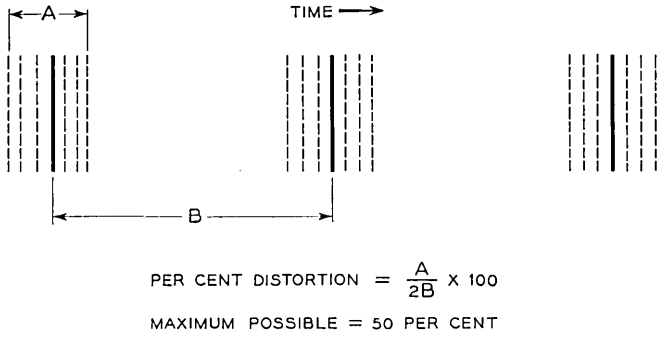


Fig. 4 — Measurement of peak jitter in terms of per cent peak distortion.

The following paragraphs discuss the effect of present telephone facility transmission characteristics on binary data signals and summarize the results of some of the basic transmission measurements that were made during the field testing program.

IV. BASIC TRANSMISSION CHARACTERISTICS

Considerable progress is being made in establishing the relationships between data system performance and the transmission parameters characterizing the telephone network. A general presentation is not within the scope of this paper and, indeed, would be premature at this time. However, the field sampling of such channel characteristics as amplitude and delay distortion, noise and net loss, together with a limited theoretical study, have made possible a first-order description of the data transmission capabilities of the telephone plant. The measured data are presented herein on a statistical basis.

4.1 *General*

Nyquist[25] theorizes that a channel should be capable of transmitting binary digital information at a rate numerically equal to twice the channel bandwidth, e.g., 6000 bits per second, assuming a bandwidth of 3000 cps. This requires a channel having flat loss and no delay distortion within the passband and infinite loss outside — conditions not met by the switched telephone network. Telephone bandwidths have been designed to accommodate speech frequencies from about 300 cps to about 3300 cps. It is therefore necessary to translate the data signal spectrum into this nominal passband by such means as the FM digital subset. If the resulting sidebands are transmitted symmetrically, the allowable bit speed is reduced by one-half.

Another factor limiting data speeds involves an effect of nonlinear distortion. It is frequently called "Kendall effect"[26] because its occurrence was first predicted by B. W. Kendall in connection with studies of telephoto transmission. Nonlinear distortion results in second-order modulation products that may fall within the baseband spectrum of the data signal. If this overlaps the line signal, distortion will result. Therefore the lower portions of the telephone band cannot always be used efficiently and the frequency space available for the data signal is reduced.

Practically speaking, then, data speeds of binary signals on the switched telephone network are certainly less than 3000 bits per second, although higher speeds may be achieved by other than binary systems. For a given speed, the rate at which errors occur will depend on the method of modulation and transmission characteristics of the channel. The basic transmission phenomena of interest are:

i. *effective channel bandwidth*, characterized by the attenuation and delay distortion parameters of the telephone network, which imposes an upper bound on transmission speed and reduces the noise margin to error generation;

ii. *circuit net loss*, which affects signal-to-noise margins and hence margin to error;

iii. *noise*.

4.2 *Transmission Characteristics of the Telephone Plant*

The characteristics described herein represent the cumulative effects of the different transmission systems and switching equipment required to complete each particular connection. Consider initially the effects of individual transmission and switching facilities.

The attenuation of typical nonloaded wire pairs is proportional to the square root of frequency within the voice band and only for short lengths is this distortion across the band tolerable. Cable pairs longer than about three miles are loaded at uniform intervals with inductance to reduce attenuation frequency distortion and the over-all loss. With this added inductance, the line looks like a low-pass filter and exhibits a cutoff. Fig. 5 is a plot of the attenuation per mile for 22-gage pairs, both loaded and nonloaded, as a function of frequency normalized to the loaded pair cutoff frequency. The cutoff of the loaded facility also introduces additional phase or delay distortion over the nonloaded pair, as shown in Fig. 6.

Carrier systems exhibit cutoffs both at high and low frequencies, as shown in Fig. 7. For clarity, the reference flat loss values are displaced

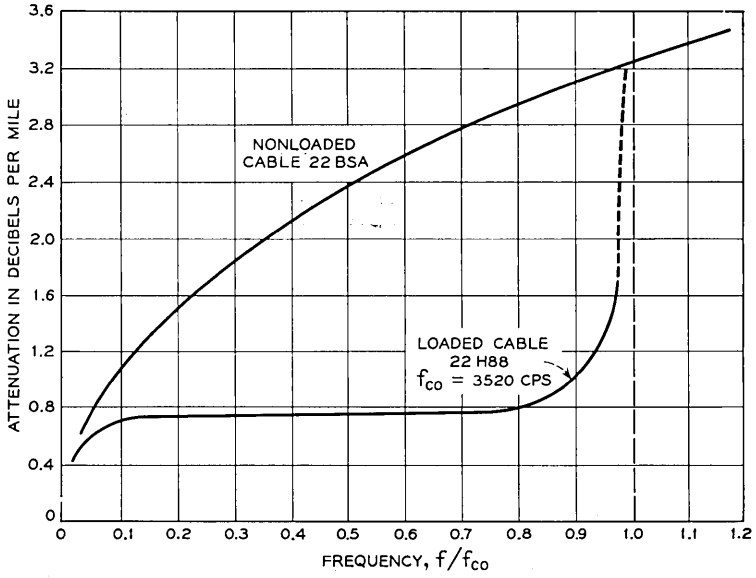


Fig. 5 — Attenuation characteristics, nonloaded and loaded cable.

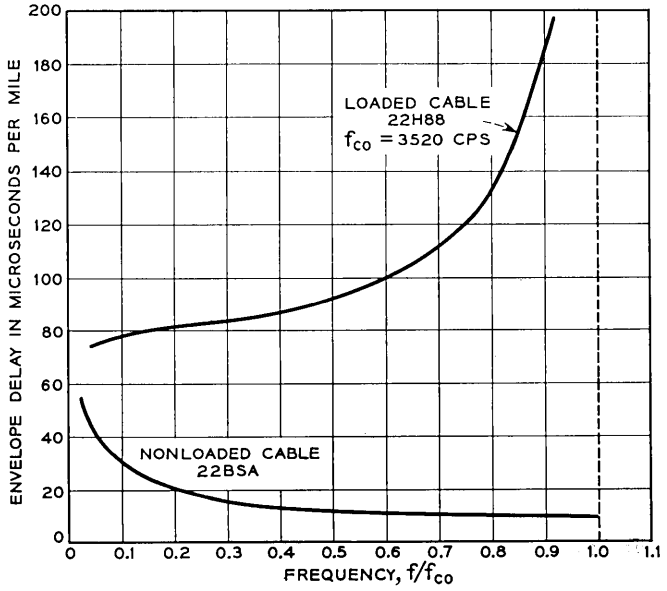


Fig. 6 — Envelope delay characteristics, nonloaded and loaded cable.

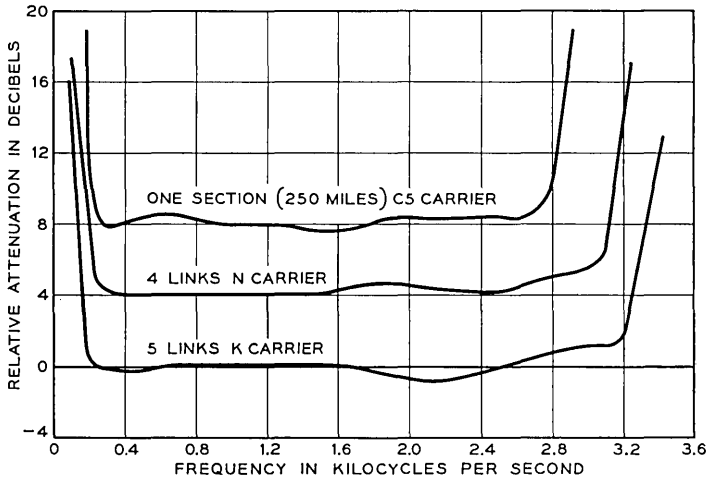


Fig. 7 — Representative attenuation characteristics, carrier systems.

vertically. Typical delay distortion characteristics for these systems are shown in Fig. 8.

Because of multiple connections and cabling runs within switching offices, shunt capacitance is added to a switched connection. This, of course, has the greatest effect at the upper end of the voice band on both attenuation and delay characteristics. Associated with switching points are the repeating coils, series capacitors and shunt inductors used for

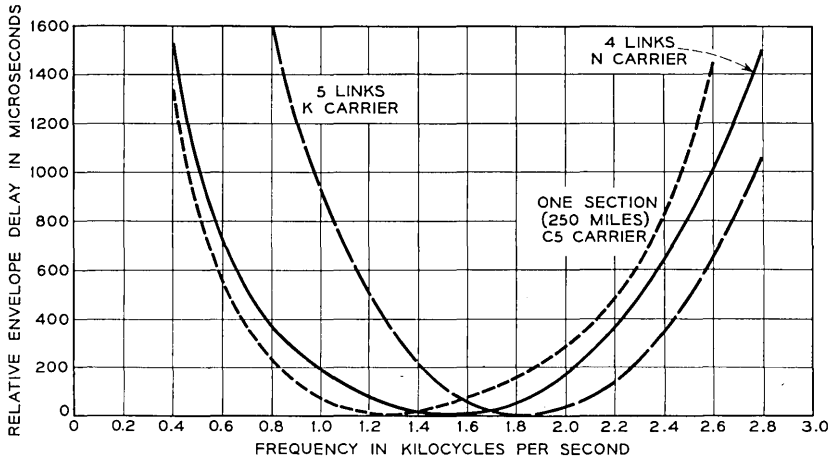


Fig. 8 — Relative envelope delay, carrier systems.

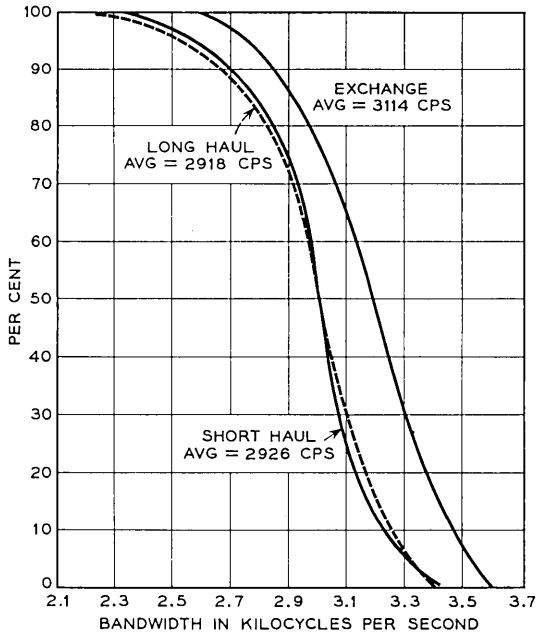


Fig. 9 — Cumulative distributions of 20-db bandwidths, showing percentage of circuits having bandwidths greater than that shown on abscissa.

signaling and supervisory purposes. These have their greatest effect at low frequencies. Therefore, even if the transmission facilities are voice-frequency wire lines, switched connections will show lower-end cutoffs.

At switching points a considerable amount of impulse noise is generated by the switches themselves, relays, dialing pulses, and the like. This noise is coupled in varying degrees, either directly or as cable crosstalk between pairs, to all channels switched by the office.

4.2.1 *Effective Channel Bandwidth*

It is convenient to consider attenuation and envelope delay distortions as occurring between two cutoff frequencies at which signal frequency components will be so severely attenuated by the transmission medium as to be relatively insignificant. For the purpose of this presentation, a 20-db bandwidth is defined as the interval between those frequencies at which the circuit loss is 20 db greater than the minimum loss of the circuit. Accordingly, Fig. 9 is a plot of cumulative distributions of 20-db bandwidths for the three classes of calls, showing the per-

centage of calls having bandwidths greater than the corresponding abscissa value. Note that the average 20-db bandwidth is on the order of 3000 cps. However, distortions to be described within this band are such that considerably less than 3000 cps may be usable for data transmission purposes.

4.2.2 *Attenuation Distortion*

A careful examination of all the characteristics taken during the field measurement program has revealed that, in general, the relative attenuation characteristics assume the form shown in Fig. 10. That is, the circuit loss rises rapidly below f_1 cps and above f_3 cps, is relatively flat from f_1 cps to f_2 cps and rises linearly with frequency from f_2 cps to f_3 cps. These average frequencies and loss roll-offs are described in Table III.

For exchange calls, sharp lower roll-offs are not to be expected on the average, since many such connections are short, use voice facilities and cut off only because of the signaling and supervisory networks. Some longer calls use carrier facilities showing a sharper cutoff. Long distance calls, in particular, use single carrier systems extensively so that the average lower roll-off is fairly sharp.

The upper end roll-offs are much sharper for all classes of calls because of the combined effects of carrier systems and inductively loaded cable pairs.

Since data signals in most cases tend to be placed in the band from

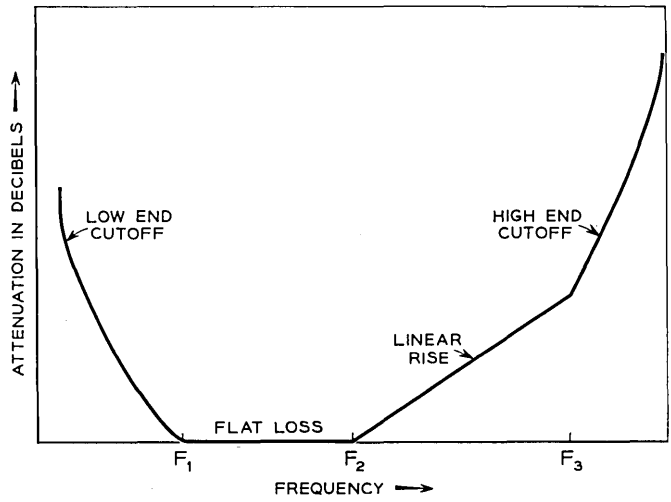


Fig. 10 — Relative attenuation characteristic of telephone circuits.

TABLE III

	Roll-Off Below f_1, db per octave	f_1, cps	f_2, cps	f_3, cps	Roll-Off Above f_3, db per octave
Exchange	15	240	1100	3000	80
Long distance					
Short-haul	24	300	1075	2950	90
Long-haul	27	280	1150	2850	80

1000 to 2600 cps, it is particularly desirable to describe the linear portion of the relative loss curve between these two frequencies. Fig. 11 is a plot of cumulative distributions for the loss differences between 1000 and 2600 cps for the three classes of calls. Note that, on the average, this difference is about 8 db but that, in about 5 per cent of the cases, 15 db is exceeded. In general, long distance connections show greater slopes, as a result, in part, of the shunt capacitance added by the switching points. Exchange calls usually are switched only twice, whereas long distance calls may be switched at four or more points.

With transmission at 1200 bits per second with the FM digital subset, it was found advantageous to use an attenuation equalizer designed to

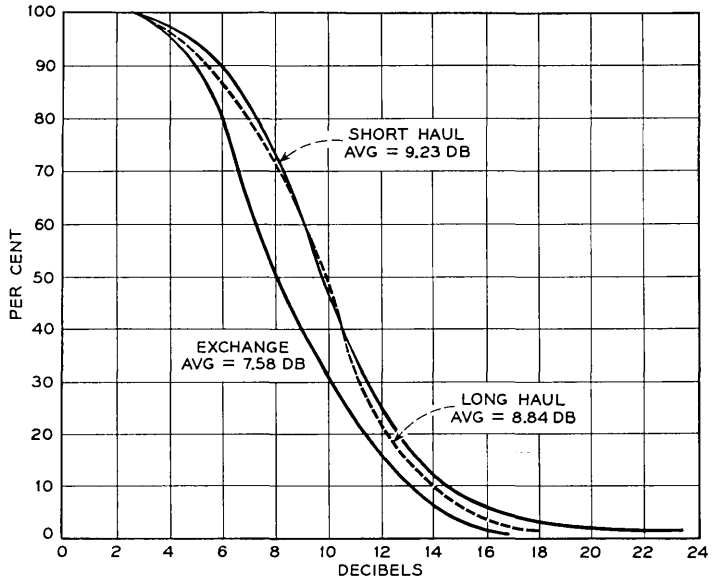


Fig. 11 — Difference in decibels between 1 kc and 2.6 kc — percentage of circuits having decibel difference greater than that shown on abscissa.

compensate for a 4-db slope between 1100 and 2100 cps, which are the mark and space data frequencies respectively. Between 1000 and 2600 cps this network equalized a loss slope of about 7 db.

4.2.3 *Envelope Delay Distortion*

The ear is relatively insensitive to minor phase distortions, so that the telephone message plant, designed for speech transmission, has not required the extremely low distortions demanded by data transmission. Since there is no reason to assume that the telephone network is minimum phase, knowledge of attenuation characteristics must be supplemented by a characterization of the phase variations. It is most practical to measure the derivative of phase with respect to frequency, $d(\theta)/d\omega$, which has the dimension of time and is referred to as *envelope delay.*

Curves of envelope delay versus frequency tend to be concave upward as a result of the upper and lower cutoffs of the telephone network. Average envelope delay characteristics are plotted in Fig. 12 for the three classes of calls, with the minimum envelope delays normalized to zero. They were derived by drawing smooth curves through the following five points: the average frequency of minimum delay (FMD), the average upper and lower frequencies at which the envelope delay is 1.0 millisecond greater than the minimum, and the average upper and lower 0.5-millisecond frequencies.

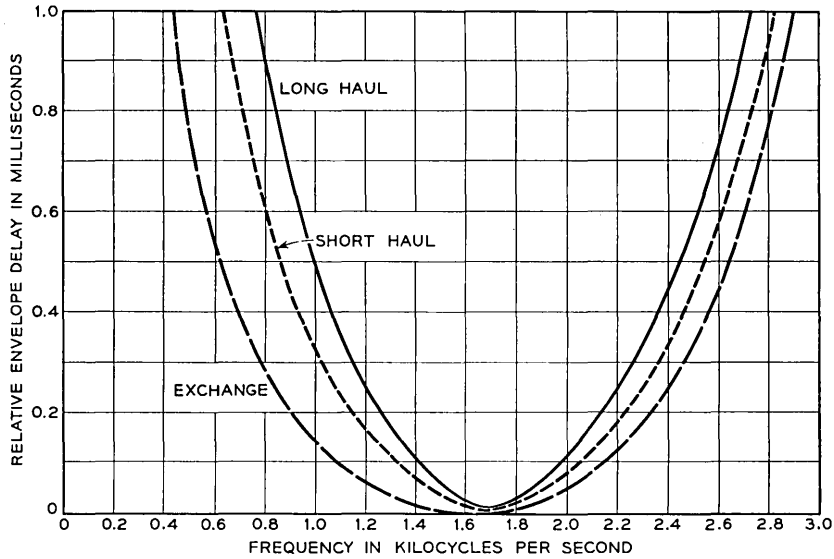


Fig. 12 — Average envelope delay characteristics.

Three facts are noteworthy: (a) the average FMD is on the order of 1700 cps; (b) distortion for exchange calls is less than for long distance calls and (c) all the curves appear to be fairly symmetrical about their respective FMD's.

It is mildly surprising that the exchange delay characteristics do not show more dissymmetry around an FMD somewhat lower than measured. Such a result is to be expected if the lower cutoff is determined by signaling and supervisory circuits. The explanation lies in the fact that almost 50 per cent of the exchange connections measured used carrier facilities with typically symmetrical delay curves. Deleting the data from calls using exchange carrier systems gives rise to the curve shown on Fig. 13, which is somewhat more representative of wire line characteristics.

Of interest are the variations from these average curves that are shown in Fig. 14. For each point used to draw the average characteristics described above, limits were found so that about 90 per cent of the measured points fell within these limits. By systematically joining respective limit points, the plots in Fig. 14 were derived. Careful sampling of the actual data confirms that approximately 90 per cent of the measured curves do fall within the shaded areas of the diagram, even though the

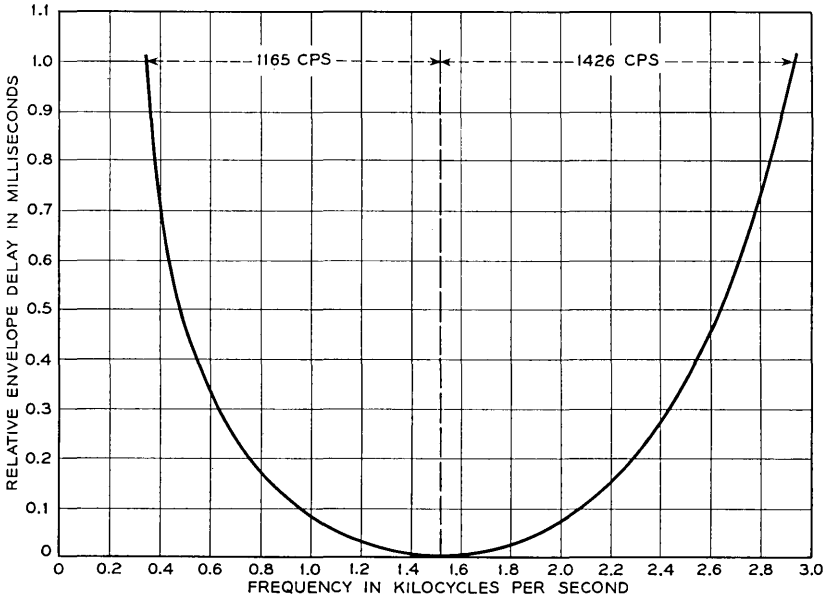


Fig. 13 — Average envelope delay characteristics for exchange calls on cable.

limit points were determined independently. Dotted lines indicate typical measured curves.

Note that all curves are tangent to the abscissa representing minimum delay (zero microseconds) at frequencies varying from 1200 to 2000 cps. More detailed information on the variations of this frequency of minimum delay is shown in Fig. 15.

Statistics on the delay distortion at the band edges are presented in Figs. 16 and 17 in terms of 0.5-millisecond and 1.0-millisecond "bandwidths." Delay bandwidth is here defined as the difference between those frequencies at which the envelope delay distortion is 0.5 or 1.0 millisecond.

A comparison is made in Fig. 18 of the measured delay characteristics with the compromise delay equalizer used during the tests with trans-

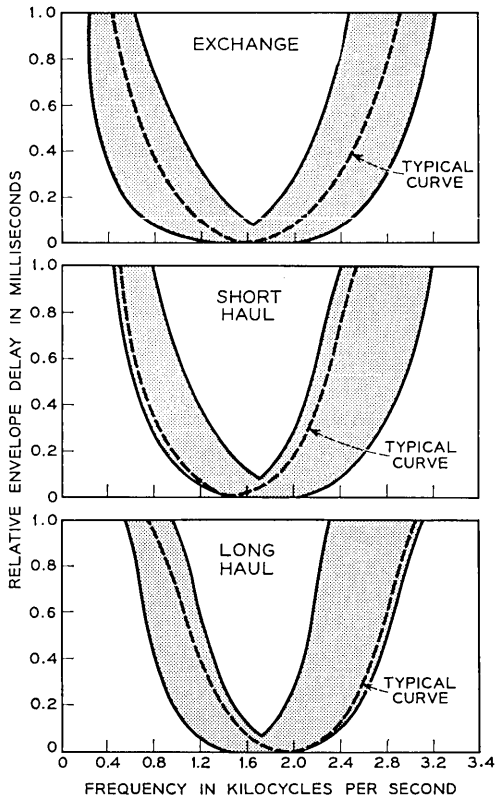


Fig. 14—Envelope delay distortion—locus of 90 per cent of circuit characteristics.

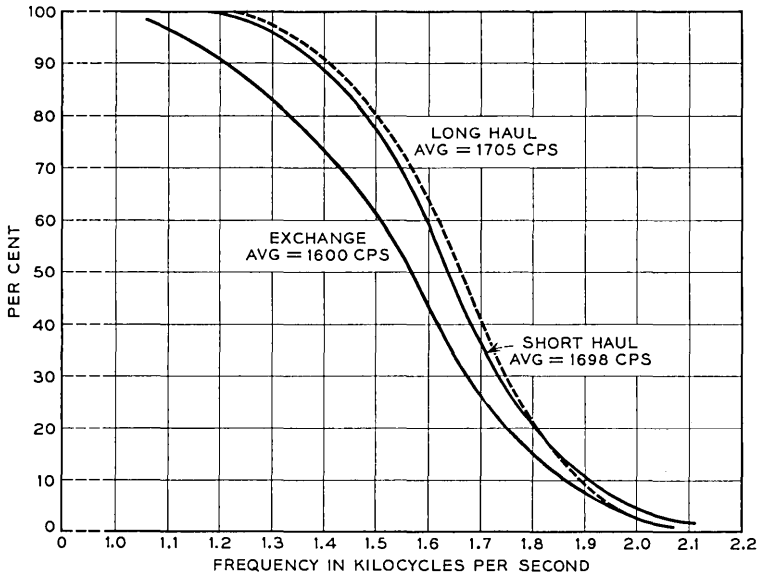


Fig. 15 — Frequency of minimum envelope delay — percentage of circuits having frequency of minimum delay greater than that shown on abscissa.

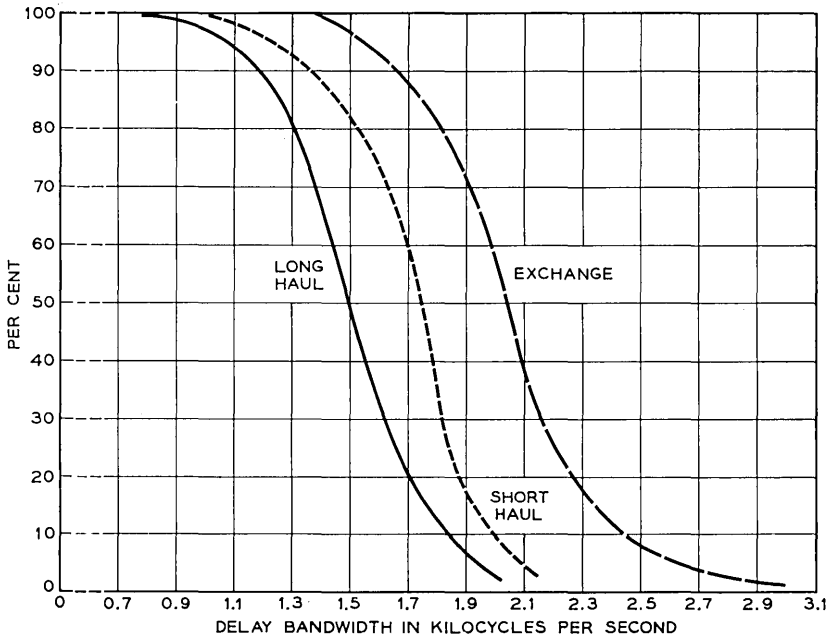


Fig. 16 — Percentage of circuits with 0.5-millisecond delay bandwidth greater than that shown on abscissa.

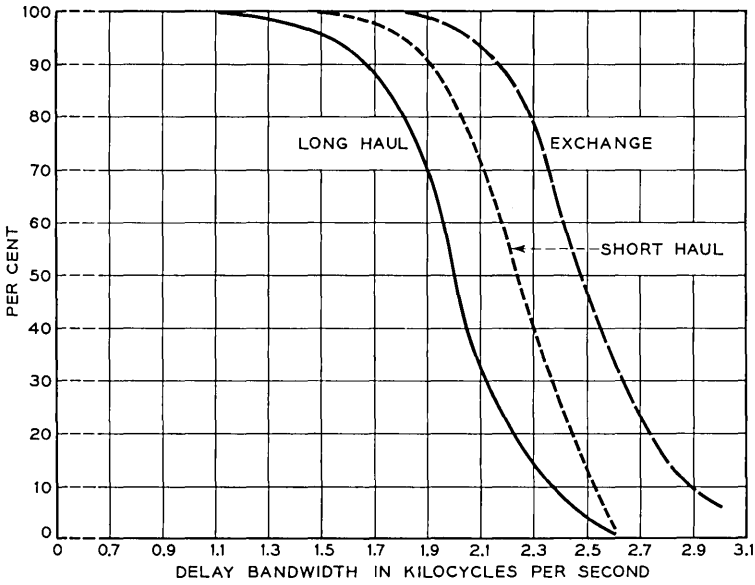


Fig. 17 — Percentage of circuits with 1-millisecond delay bandwidth greater than that shown on abscissa.

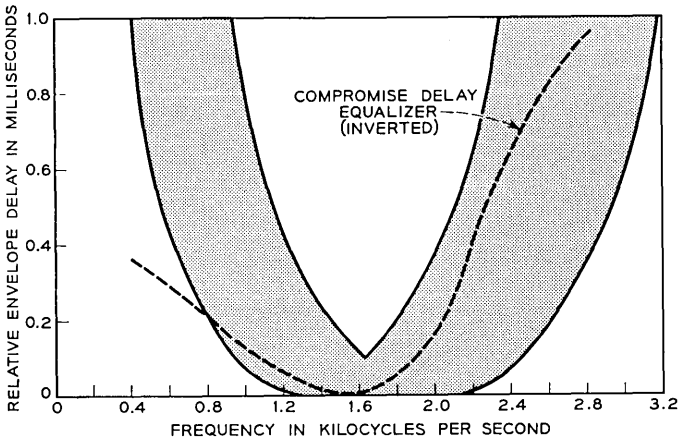


Fig. 18 — Envelope delay distortion characteristic for all calls — locus of 90 per cent of all circuit characteristics.

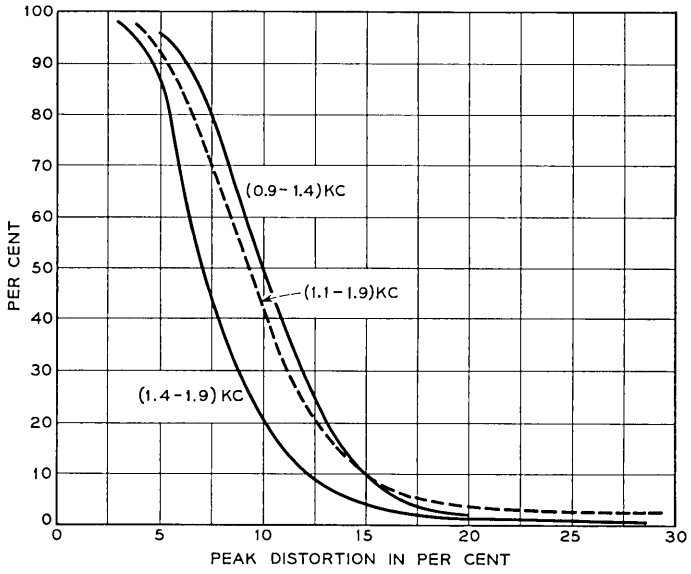


Fig. 19 — Peak distortion for 600-bit-per-second calls, no equalization — percentage of circuits having peak distortion greater than that on abscissa.

mission at 1200 bits. The inverse of the equalizer characteristic is shown superposed on a diagram indicating the locus of 90 per cent of the delay curves including all three classes of calls. Hindsight indicates that lower frequencies probably would have been better equalized on the average had the compromise favored those circuits utilizing carrier facilities.

The combined effect of amplitude and delay distortion on the FM digital subset shows up as jitter on the transitions of the demodulated signal and can be described in terms of peak distortion (repetitive jitter). Peak distortions of less than 20 per cent are considered quite acceptable. Fig. 19 shows that more than 99 per cent of the calls met this 20 per cent limit while transmitting at 600 bits at mark-space frequencies of 1400–1900 cps. Although the percentage of calls exceeding the limit did not vary widely for the three frequency pairs used, the over-all distribution for the 1400–1900 cps was considerably better. This was probably due to a better match of the resultant data spectrum to the average envelope delay characteristic. A correlation of peak distortion and error performance showed that the error statistics would not have been significantly changed if the 1400–1900 cps frequency pair had been used in this test.

Upon changing to a speed of 1200 bits per second, the measured peak

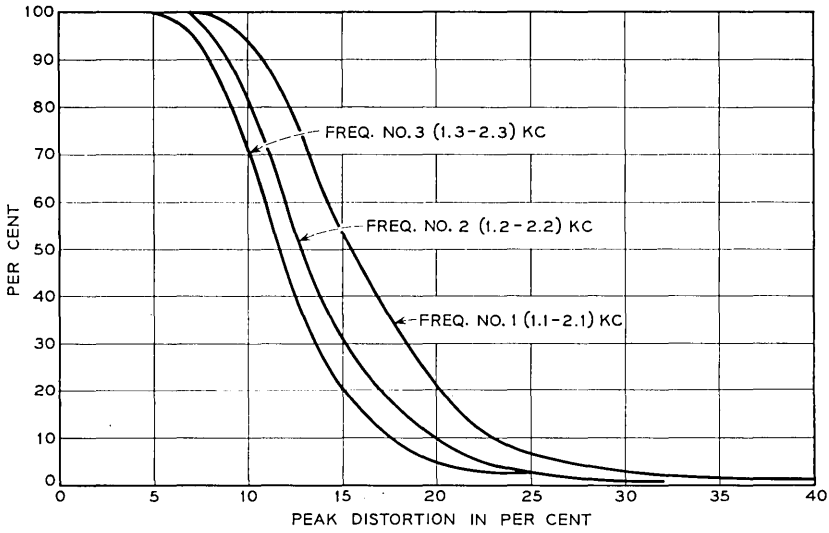


Fig. 20 — Peak distortion for 1200-bit-per-second calls with equalization (delay plus attenuation) — percentage of circuits having peak distortion greater than that shown on abscissa.

distortion, without either attenuation or delay equalization, was beyond practical limits. After the compromise equalizer was inserted, the jitter was considerably improved. Fig. 20 plots the cumulative distributions of peak distortion for the three pairs of frequencies used. Note that, for pair #1 (1100–2100 cps), over 20 per cent of all calls exceeded the 20 per cent limit, whereas for pairs #2 (1200–2200 cps) and #3 (1300–2300 cps) less than 10 per cent of the calls exceeded the limit. Pair #3 shows the lowest distortion for two possible reasons: (a) these frequencies best match the average compromise equalized connection and (b) the number of carrier cycles per signal element is greatest for this pair.

4.2.4 *Ripples in Distortion Characteristics*

The attenuation and delay characteristics described herein were derived by discounting ripples in the raw measured data. These ripples, mainly the result of echos (reflected energy), can be appreciable and must not be ignored. The source of echoes is varied and includes all points of impedance mismatch in bilateral circuits, and hybrid unbalance at the junction of two- and four-wire circuits. If multiple echoes exist, ripples in the attenuation and delay characteristics of the same circuit are not necessarily correlated. However, for each characteristic the ripple period on a frequency scale tends to be inversely proportional to the

electrical length of the path from the observer to the source of echo. That is, close-in echoes give rise to long sweeping ripples, while remote echoes cause fine structure ripples.

Due to increased reflected energy at the band edges, where impedances deteriorate, the ripple amplitudes tend to increase in these areas. In the main, the ripple in the amplitude characteristic is significant (i.e., greater than one to two db) only above about 2000 cps. An appreciation of the amount of ripple likely to be encountered can be gained by referring to Fig. 21, where a bar chart indicates the percentage of circuits having a maximum peak-to-peak ripple in decibels. For the most part, this maximum ripple occurs in the frequency range of 2000 to 3000 cps.

Both the amplitude and period of the ripples vary across the band, probably due to the existence of multiple echo points. Such variations are difficult if not impossible to describe statistically.

It has been pointed out[27] that close-in echoes result in ripples in transmission characteristics that can be equalized, whereas remote echoes cannot. The reason for this is that an individual transmitted pulse will be affected mainly by its own echo if the source of the echo is close in, and can be equalized to eliminate this distortion. Remote echoes tend to affect subsequent transmitted pulses, and the effect is random for an information bearing train. Hence ripple equalization will not be effective in general.

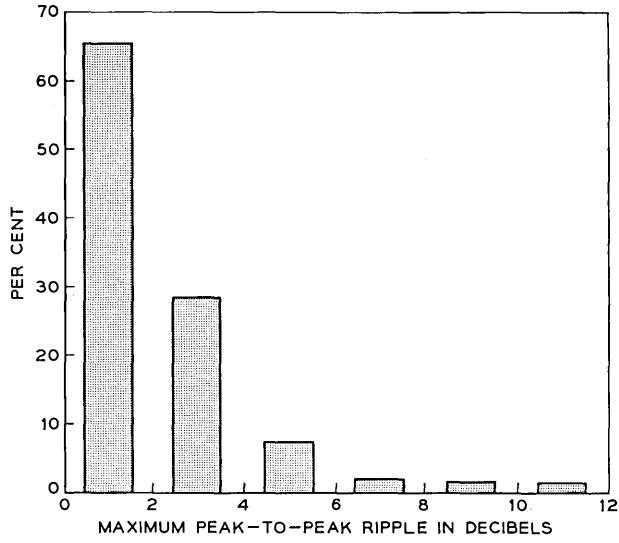


Fig. 21 — Percentage of circuits with maximum peak-to-peak amplitude ripple.

4.2.5 *Circuit Net Loss*

It is common practice to specify the net loss of telephone circuits at 1000 cps, but the actual loss for a complex frequency signal may be somewhat different, depending on the attenuation frequency characteristic of the circuit. An example of this difference will be given. Consider first the cumulative distribution of 1000-cps circuit net loss (CNL) for the three classes of calls in Fig. 22. Note that, on the average, exchange calls are a few decibels better than long distance calls. This is to be expected, since loss tends to be a function of the physical length of the connections. Since long distance connections can be thousands of miles longer than exchange calls, it is gratifying to note that this relatively small difference in loss between the two types has been achieved in practice.

Consider the loss experienced by the FM digital subset signal operating at 1200 bits per second. For mark-space frequencies of 1100–2100 cps the apparent carrier frequency is 1600 cps. Reference to Section 4.2.2 shows that the average loss at 1600 cps is about 3 db above the 1000-cps value. Taking into account the entire average attenuation characteristic across this data band — 500 to 2700 cps — an excess loss

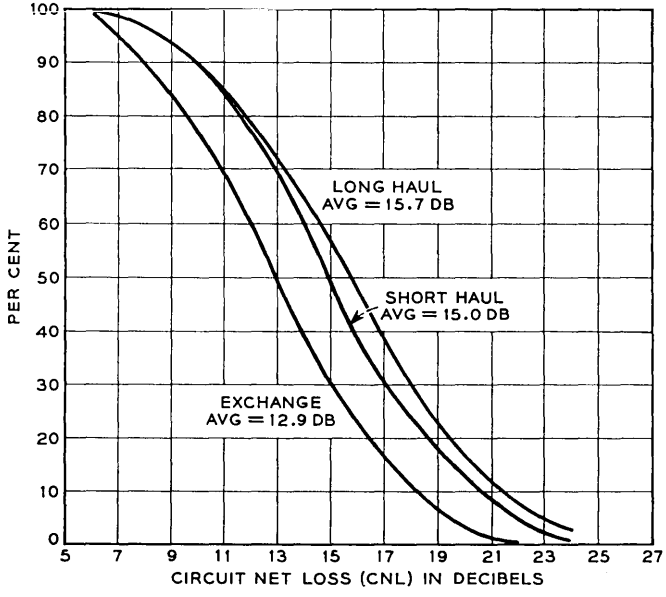


Fig. 22 — Percentage of circuits with 1000 cps net loss greater than that shown on abscissa.

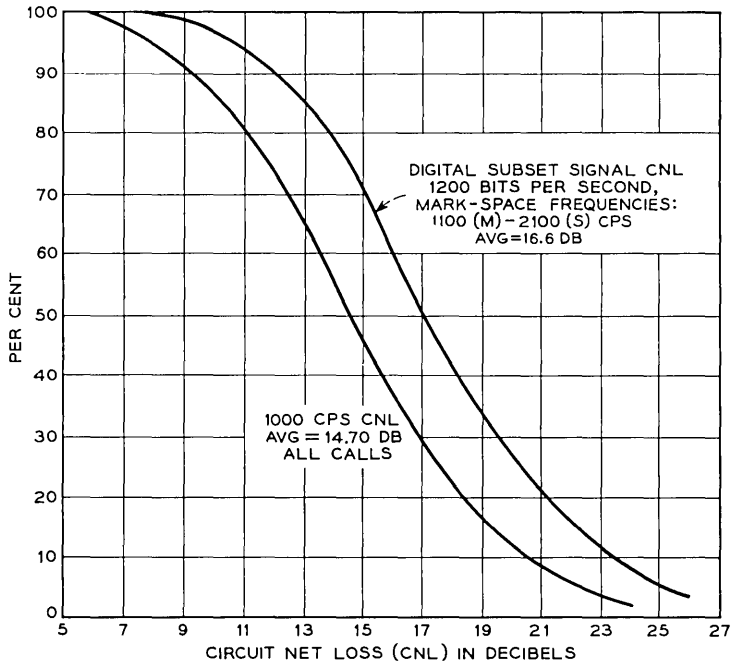


Fig. 23 — Percentage of circuits with circuit net losses greater than that shown on abscissa.

of less than 3 db over 1000-cps CNL is to be expected. Fig. 23 is a plot in the cumulative sense of the 1000-cps CNL and the data signal net loss for all calls. Note that the average data signal loss is only 2 db greater than the average 1000-cps CNL, confirming the prediction.

4.2.6 *Noise*

Two types of noise are of interest in the telephone plant: (a) steady line noise and (b) impulse noise. Steady noise is important for its interfering effect on speech transmission. Impulse noise, characterized by relatively high peaks of short duration pulses, has the greatest effect on the transmission of pulses.

To see that, in general, steady noise is not of great importance in pulse transmission, consider its cumulative distribution in Fig. 24. Note that only about 1 per cent of all the calls exceed noise values of 40 dba. This is equivalent to an average of about −42 dbm of white noise in a 3-kc band. Referring again to Fig. 23, note that in only about 5 per cent of the calls did the −6-dbm data signal exceed losses of 26 db for a

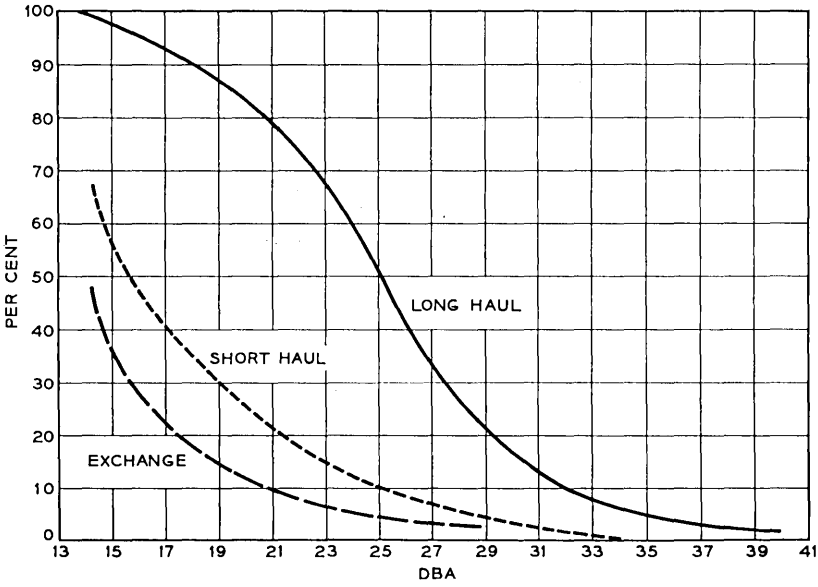


Fig. 24 — Percentage of circuits with F1A noise greater than that shown on abscissa.

received level of −32 dbm. Even without determining the degree of interdependence of the two distributions, it is apparent that very few calls exhibited less than a 10-db average signal-to-noise ratio.

Impulse noise, on the other hand, frequently has peaks comparable to the received data signal level. The incidence of impulse noise tends to follow the traffic fluctuations in the switched network. That is, busy periods generate considerably more impulse bits than do quiet periods. In fact, in the field test about 40 per cent of the calls failed to show any impulse counts regardless of the measured level. This is to be expected, since calls were placed at random during both the busy and quiet periods of the day.

The average number of counts of impulse noise above given slice levels for 15-minute measurement periods is plotted in Fig. 25. These data give a general indication of impulse noise conditions within the message plant even though they do not correlate well with error rates on the same calls. In many instances in which errors occurred, no impulse noise was measured, and vice versa. As a result, the correlation of impulse noise and error generation was poor. Drop-outs and interruptions that do not show up as impulse noise counts limit the usefulness of noise measure-

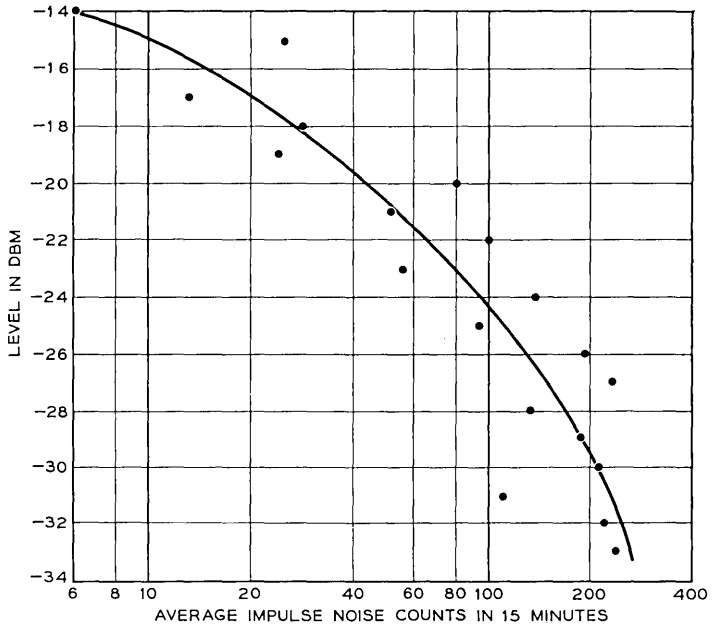


Fig. 25 — Average impulse noise counts in 15 minutes above level shown on ordinate.

ments in predicting error performance on the switched message network.

The error performance actually experienced during the field testing program is described and discussed in the following paragraphs.

V. ERROR RATES

The primary purpose of the investigation described herein was to provide statistics of the error rate and the time distribution of errors at bit rates that have a *high probability of success for transmission over any facility between any two telephone stations in the United States.* There would be little value in designing data systems that had such high bit rates that they would work on only half of the circuits encountered. We are interested in the performance at data rates where satisfactory transmission can be achieved on practically all of the connections.

Success in data communication does not mean that the communication must be completely error-free. The terms "successful" or "satisfactory" are as difficult to define for data as they are for speech com-

munication. Very few people carry out a conversation — even in the same room, much less over a communication facility — without the necessity of repetition or the deliberate insertion of redundancy in vital parts of the message because of distraction on the part of the listener. This distraction is often extraneous noise or reverberation in the room. On a communication facility, this distraction might be noise on the circuit or distortion of the signal resulting in nearly the same effect as noise on the circuit. Therefore, a request is made on the part of the listener to repeat. The communication is usually considered successful or satisfactory until the point is reached where the distraction becomes high enough to require an annoying amount of requests for repetition. This will vary with the articulation and modulation of the speaker, the text of the message and the patience of the communicators, as well as the transmission characteristics of the circuit. For Data-Phone equipment the same philosophy applies. The evaluation of performance is more easily defined in data because of the binary nature of the information and because the so-called distraction results in a recognizable error. However, redundancy either in the form of repetition or check digits, or both, can be used to improve the accuracy, and the need for it is a function of the same variables. By the proper use of redundancy it is possible to achieve any desired degree of accuracy.

In order to obtain a better understanding of error rate as a function of transmitting level, some measurements were made at a number of levels. The bulk of the data, which includes the distribution of errors as a function of time, was collected with a transmission level of -6 dbm at the sending station.

The method of recording clock pulses and error pulses on magnetic tape, as previously described, permits a computer to count the number of good bits between error bits and present the distribution of errors. This distribution is obtained in a printed output similar to that shown below, and is also available on cards and on magnetic tape, which can then be used for later evaluation of various types of error-control schemes or for more detailed analysis of error bursts:

Zeros	*Ones*
25,226	1
222,866	1
14,692	6
8,971	1

The first column designated "zeros" is the number of good bits between errors. The second column designated "ones" is the number of

errors. Thus, this printed output is interpreted as follows: 25,226 good bits were transmitted and then one error was encountered; then 222,866 good bits were transmitted and another error was encountered, then 14,692 good bits were transmitted and six consecutive errors were encountered, etc. This is the basic information from which various types of analysis have been made.

The particular distributions and relations presented herein have been selected on the basis of what is thought to be most significant in the planning of data communication systems. The first statistic essential in the planning or evaluation of a data system is the cumulative distribution of average error rates.

5.1 *Average Error Rates at 600 Bits Per Second*

Figs. 26 through 29 indicate the probability of obtaining a circuit that produces an average error rate better than that shown on the abscissas. These probabilities are shown for the three types of calls. It is at the receiving central office where the introduction of switching noise is most critical due to the lower level of the signal. Fig. 26 indicates that 85 per cent of the exchange calls can handle 600 bits per second with an error rate of one bit in error for every 10^5 bits or more transmitted. A slightly lower percentage, 82 per cent, of the short-haul calls performed as well, and 75 per cent of the long-haul circuits met this accuracy figure. On the

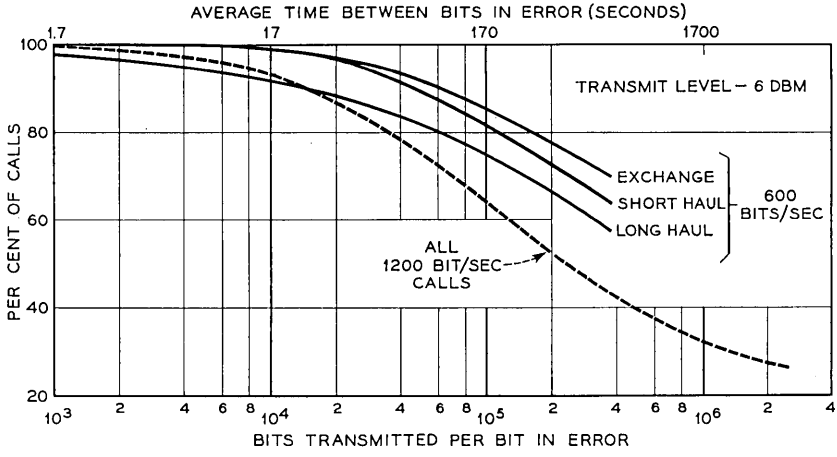


Fig. 26 — Error-rate distribution by class of call, 600 bits per second — percentage of calls with average error rate equal to or better than that shown on abscissa.

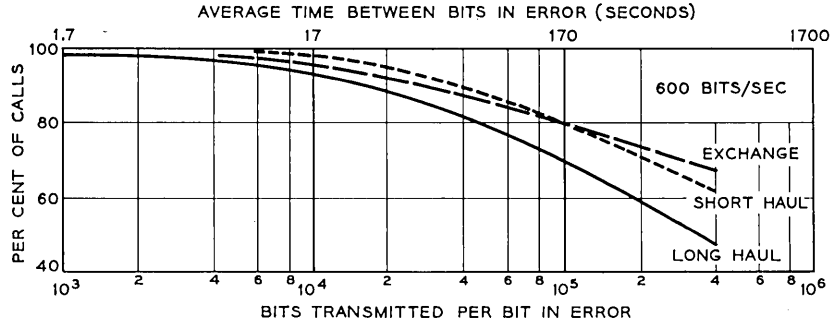


Fig. 27 — Percentage of calls having average error rate better than that shown on abscissa — crossbar office receiving, receive level −25 db.

average, exchange calls have less attenuation than short-haul or long-haul calls. Since all stations are transmitting at the same levels, this means that the signal-to-noise ratio at the receiving station is greater for the exchange calls.

In order to eliminate the effect of the higher losses on the longer connections, Fig. 27 compares error rates at a common receive level of −25 dbm, the transmitting levels being adjusted so that signals of −25 dbm were received at the receiving station line. Here the exchange calls and the short-haul calls are virtually the same, but on the long-haul calls at

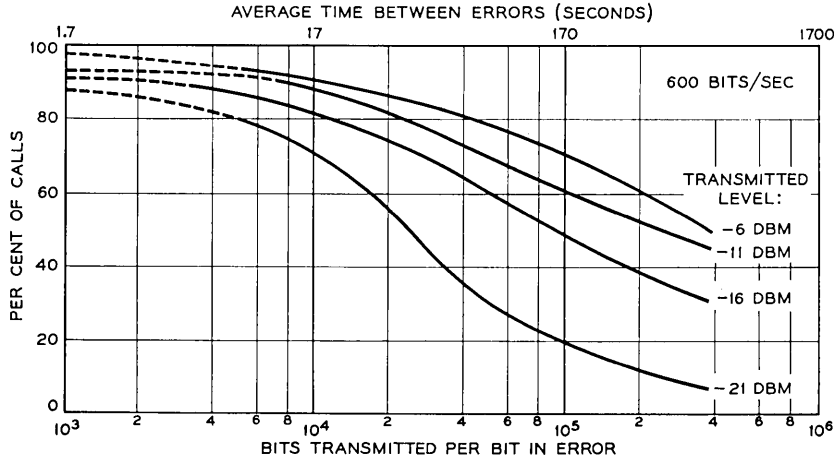


Fig. 28 — Long-haul toll calls — percentage of circuits with error rates better than that shown on abscissa as transmitting level is reduced.

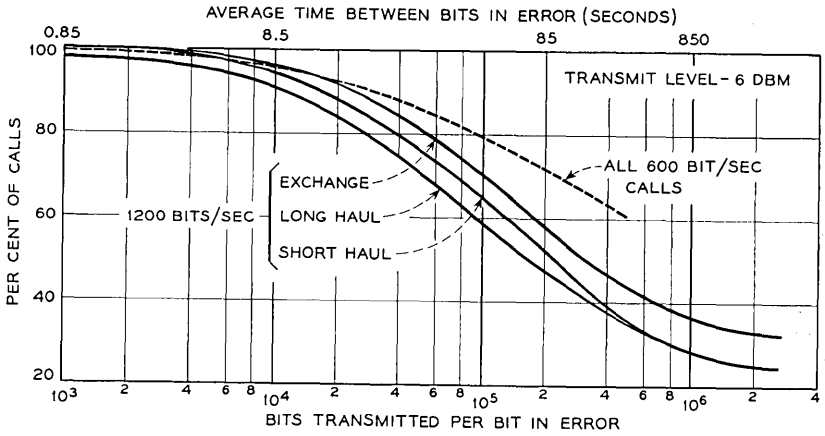


Fig. 29 — Error-rate distribution by class of call, 1200 bits per second — percentage of calls with average error rate better than that shown on abscissa.

the same receive level the performance is somewhat inferor. This should be expected, since on long circuits there is a greater probability of exposure to noise and interference.

Fig. 28 indicates the change in error rate distribution as the transmitting level is reduced in 5 db steps. Note that if the transmitting level is lowered from −6 dbm to −11 dbm the error rate is approximately doubled.

5.2 *Average Error Rate at 1200 Bits Per Second*

The curves in Fig. 29 are plots of the results at 1200 bits per second for the various types of calls. They are shown on the same axis with the curve of all calls at 600 bits per second. Note that, for a given percentage of the circuits, the error rate at 1200 bits per second is two or three times greater than that for 600 bits per second. In other words, 70 per cent of the calls at 600 bits per second will produce an error rate better than about one error in 200,000 bits, but at 1200 bits per second 70 per cent of the calls will produce an error rate better than one in 70,000 bits.

VI. ERROR DISTRIBUTIONS AND ERROR-CONTROL EVALUATION

A data transmission system should be designed to provide for the optimum useful bits of information with the minimum cost. Cost includes the cost of data equipment such as the data originating and receiving equipment and the cost of providing a communication channel,

as well as that of the modulators and the demodulators. In many cases, this resolves into the design for optimum line efficiency for a specified accuracy objective. Many studies have been made to relate this efficiency in terms of various error correcting and error detecting methods. Wood[28] derives optimum block lengths for retransmission methods, and Brown and Meyers[29] describe the efficiency of various error-control systems, including forward-acting codes and retransmission methods. However, in all these evaluations the probability of errors and the distribution of errors in time are fundamental in arriving at the proper solution. The selection of optimum codes and optimum block lengths in error-control schemes is a complex subject. The information contained in the statistics herein is provided to aid in the derivation of better control systems. However, the over-all system concept for data transmission, including error control, should be cognizant of the following considerations:

i. How serious is an error that is produced? Is error control necessary in view of the accuracy of the origin of the data or the final disposition of the data?

ii. Is the relationship of the line transmission cost to equipment cost including error control such that optimum line efficiency may not result in the most economical solution for the system?

iii. Is the format of the data such that optimum blocking must be in terms of lines of characters or numbers of cards where mechanical limitations are an important factor in the optimum arrangement?

iv. Is there storage and logic circuitry already provided in the system, such as in a computer or in buffer storage of other data machines, which can also be used for error control purposes?

The above factors are functions of the data machinery and how it is employed. In addition, the functions of the transmission medium, such as error probability, propagation time and turn-around time of echo suppressors, must be considered to resolve the optimum data transmission system. If it were not necessary to consider all these factors, then the error-control function could become a basic feature of the transmission medium. Therefore, it is not the purpose of this paper to make an evaluation of the many specific error-control methods that have been proposed, but it is desired to provide the fundamental error distributions and indicate the relative orders of magnitude of improvement that might be expected from the error-control schemes. The following curves are arranged to facilitate evaluation of optimum block lengths for retransmission methods, to evaluate error detection schemes, and to evaluate forward-acting single-error and multiple-error correction codes, including burst-correcting codes.

The error-rate distribution curves (Figs. 26 through 29) describe the probability of getting an error per number of bits transmitted. An important statistic is the probability of getting succeeding errors within various time intervals after the first error, for it is the dependency of one error on another that must be considered in error-detection or error-correction codes. If a cumulative distribution is made of the numbers shown in the previous table under the "zero" column, which represents the good bits between errors, a curve is obtained which shows the probability of getting an error as a function of time since the previous error. Figs. 30 and 31 show these distributions for 600 bits per second and 1200 bits per second, respectively. The results have been analyzed and plotted for exchange calls, short-haul calls and long-haul calls, and another curve for all calls together has been drawn.

These curves provide statistics that are useful in the planning and evaluation of error-control schemes. For example, after an error has occurred, the probability of having one or more good bits following that error before getting another error is 0.70 considering all types of calls. This means that the probability of having zero good bits, which is the

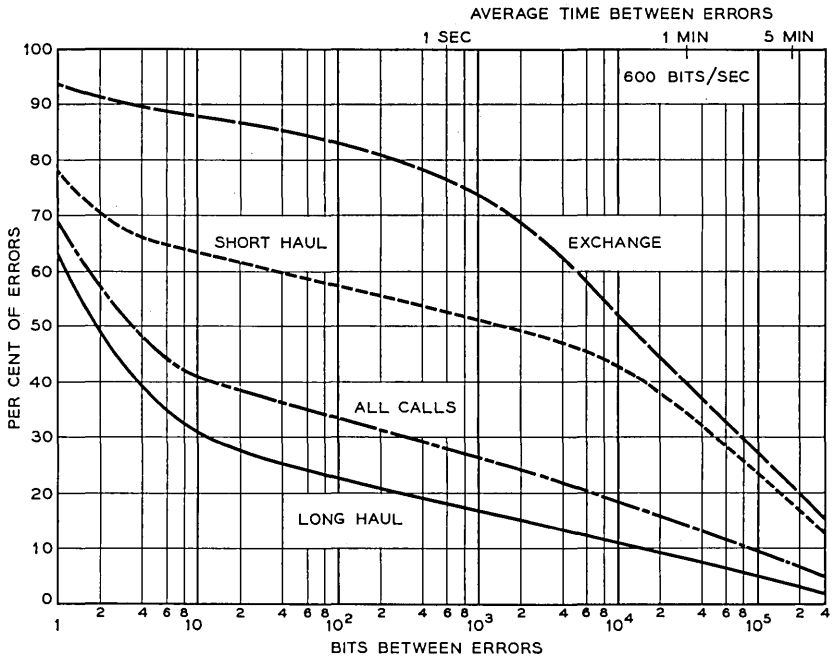


Fig. 30 — Error-free transmission time between successive errors, 600 bits per second — percentage of errors having as many as or more error-free bits between them as that shown on abscissa.

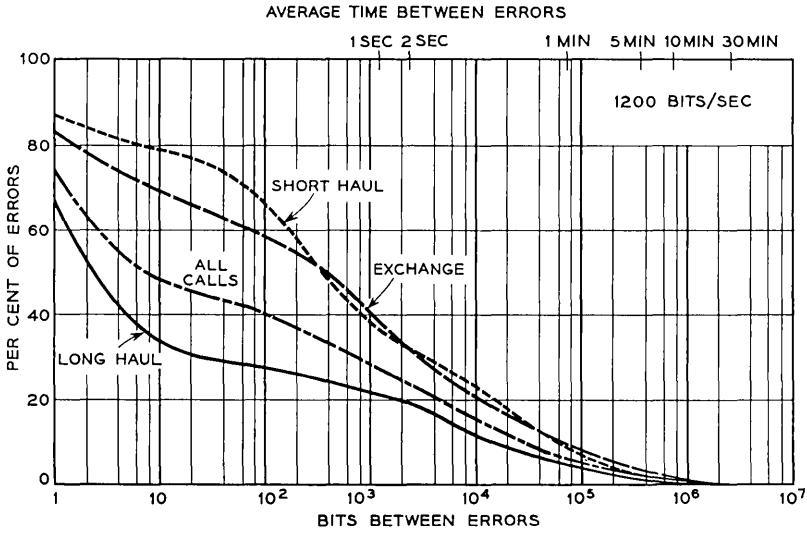


Fig. 31 — Error-free transmission time between successive errors, 1200 bits per second — percentage of errors having as many as or more error-free bits between them as that shown on abscissa.

same as having two or more consecutive errors, is 0.30. In other words, 30 per cent of all errors are immediately followed by one or more errors. For 1200 bits per second the comparable figure is 0.74, which is approximately the same as the 0.70 for 600 bits per second. If there is particular interest in eight-bit characters, for example, on long-distance calls at 600 bits per second, the curve shows that approximately 40 per cent of the erroneous characters are likely to contain single bit errors because four or more good bits will follow the erroneous bit. This assumes that, on the average, the erroneous bit is in the middle of the character. However, this means that 60 per cent of the erroneous characters will have more than one bit in error.

Each forward-acting correction code, whether it be a Hamming code,[30] a Hagelbarger code,[31] or a square matrix code, is limited in the number of errors it can correct within a given number of total bits. Also, some codes require that a period of error-free transmission exist for specific lengths of time between errors or bursts of errors in order to clear out the memory and logic of the circuitry to have it ready for the next burst of errors. The number of correctable errors in a burst and the clear-out period required is a function of the redundancy of the code and the amount of storage and logic provided in the system.

To define these bursts let us assume the sequence of good bits and error bits shown by zeros and ones below:

sequence: 0000101000000000110101000000100010000
bursts of four: |↔| |↔||↔| |↔||↔|

A *burst* is defined as a sequence of bits that starts with an error bit and extends for $N - 1$ additional bits whether they be error bits or not, where N is the length of the burst. For example, assume we are interested in burst sizes of length 4. The first bit in error and the next three bits following are considered as the burst. The succeeding burst of size 4 starts at the next error that occurs after the first burst, and so on until the entire message is analyzed by bursts of size 4 and the quantity of good bits between bursts. Thus, in the illustration above there are nine good bits between the first two bursts of 4, one good bit between the second and third burst, six good bits between the third and fourth burst, and three good bits between the fourth and fifth burst. The number of good bits between bursts, as illustrated, is counted from the last error in one burst to the beginning of the next burst. The density of the burst is the ratio of good bits in a burst to total bits. For example, in the illustration two out of a total of four bits in the first burst are in error, and the density is 50 per cent, whereas in the second burst three bits are in error, and the density is 25 per cent.

An analysis of the 600-bit-per-second transmission and the 1200-bit-per-second transmission on this basis is described in Figs. 32 through 39. A range of burst sizes from bursts of one, which facilitate the evaluation

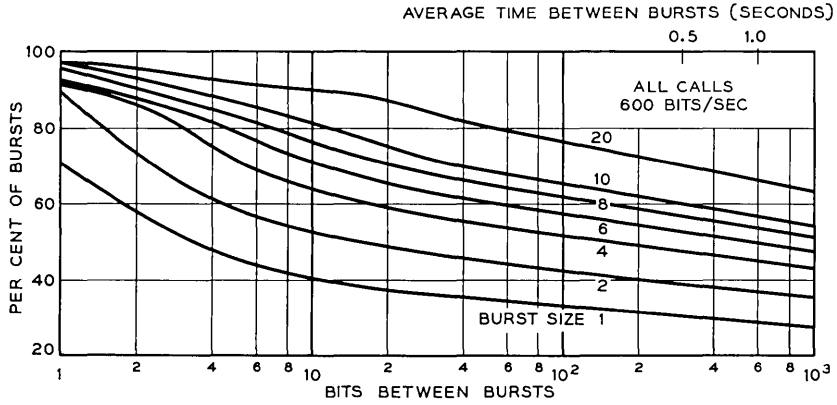


Fig. 32 — Error-free transmission time between successive bursts of various sizes, all calls, 600 bits per second — percentage of bursts having as many as or more error-free bits between them as that shown on abscissa.

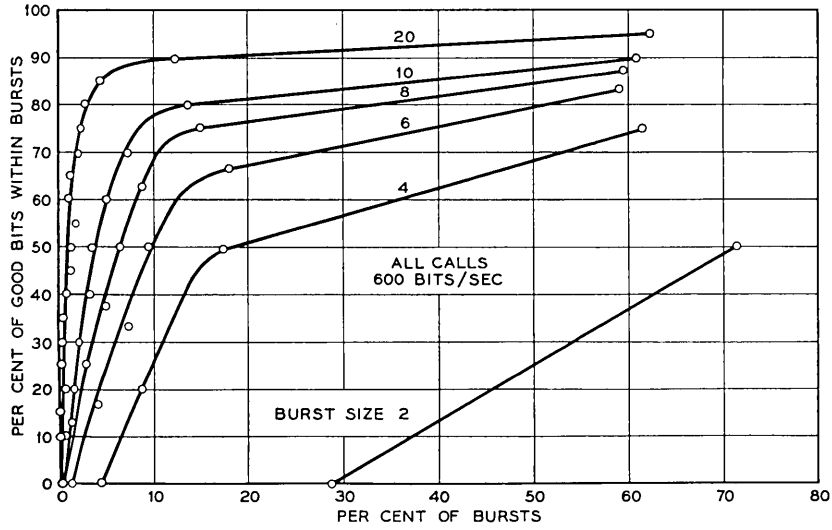


Fig. 33 — Density of good bits within bursts, all calls, 600 bits per second — percentage of good bits within bursts plotted as a function of the percentage of bursts of various sizes.

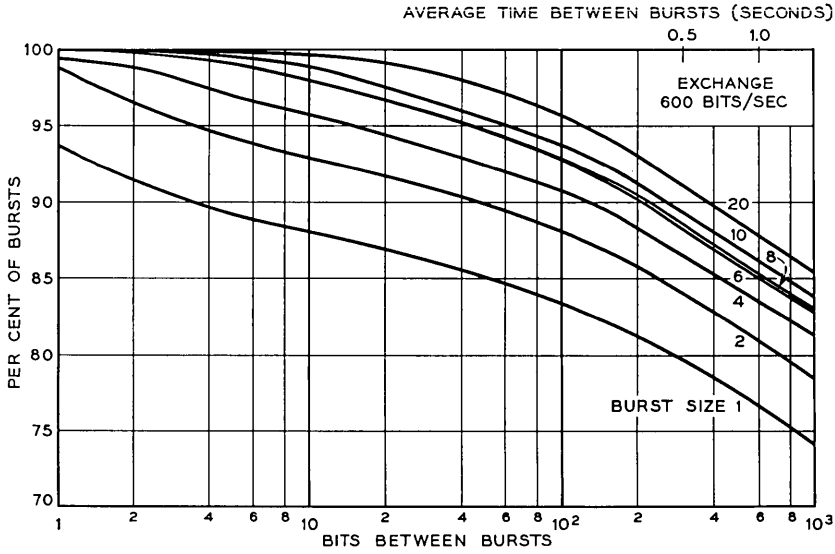


Fig. 34 — Error-free transmission time between successive bursts of various sizes, exchange calls only, 600 bits per second — percentage of bursts having as many as or more error-free bits between them as that shown on abscissa.

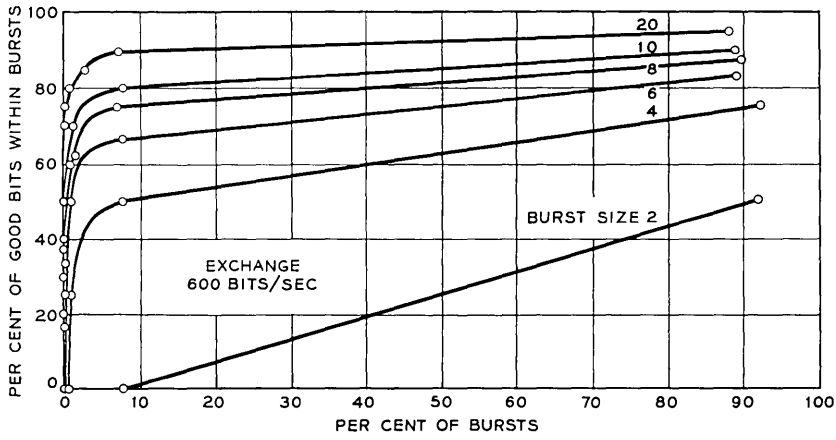


Fig. 35 — Density of good bits within bursts, exchange calls only, 1200 bits per second — percentage of good bits within bursts plotted as a function of the percentage of bursts of various sizes.

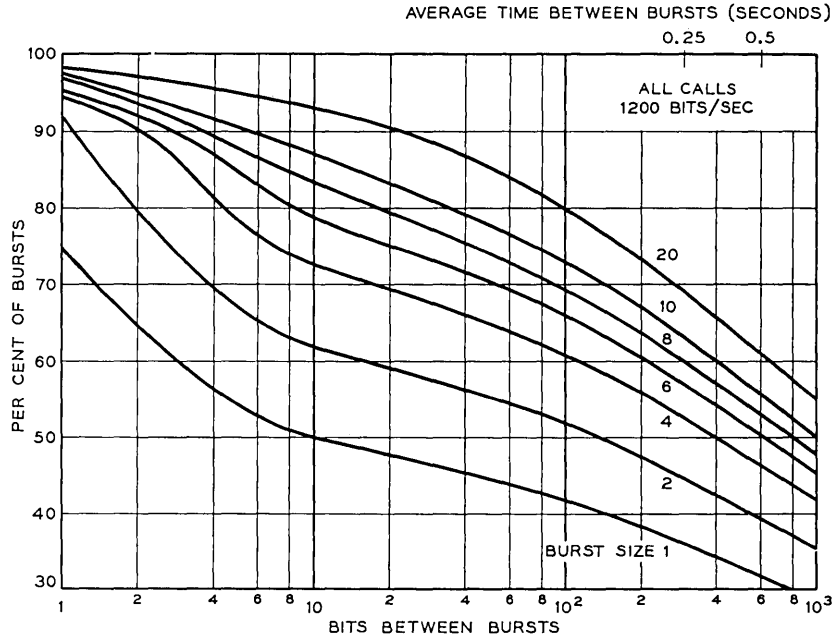


Fig. 36 — Error-free transmission time between successive burst of various sizes, all calls, 1200 bits per second — percentage of bursts having as many as or more error-free bits between them as that shown on abscissa.

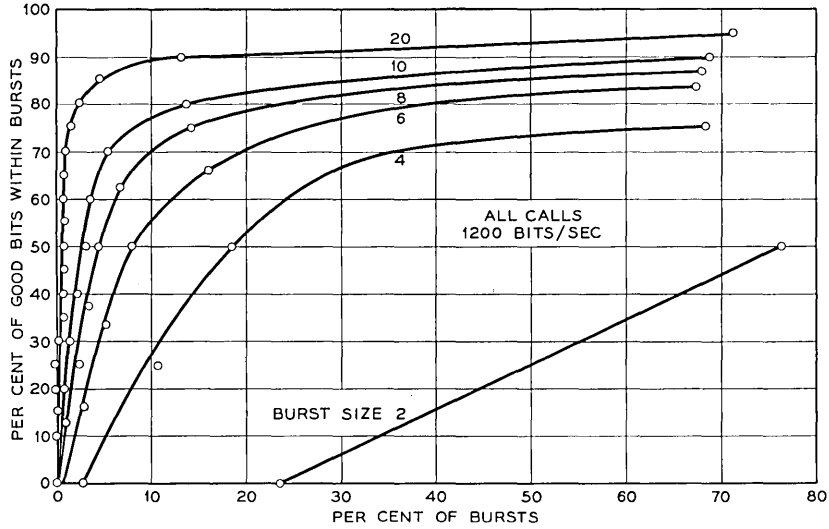


Fig. 37 — Density of good bits within bursts, all calls, 1200 bits per second — percentage of good bits within bursts plotted as a function of the percentage of bursts of various sizes.

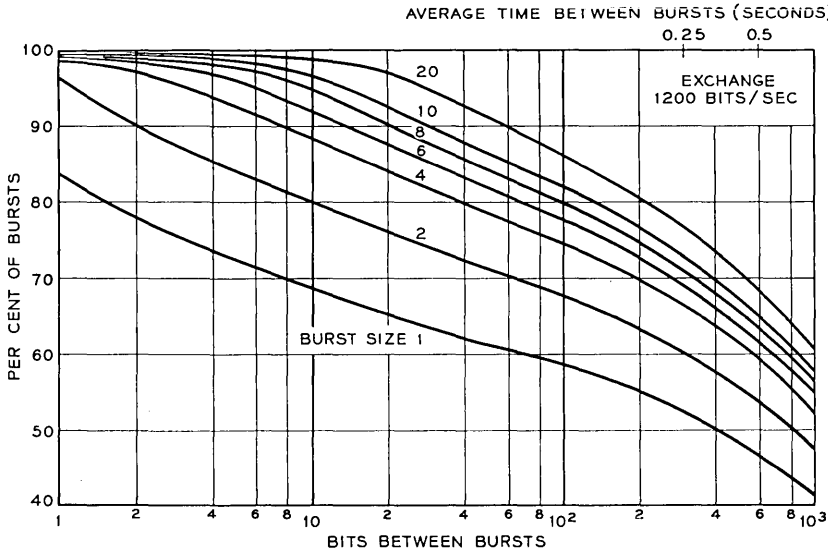


Fig. 38 — Error-free transmission time between successive bursts of various sizes, exchange calls only, 1200 bits per second — percentage of bursts having as many as or more error-free bits between them as that shown on abscissa.

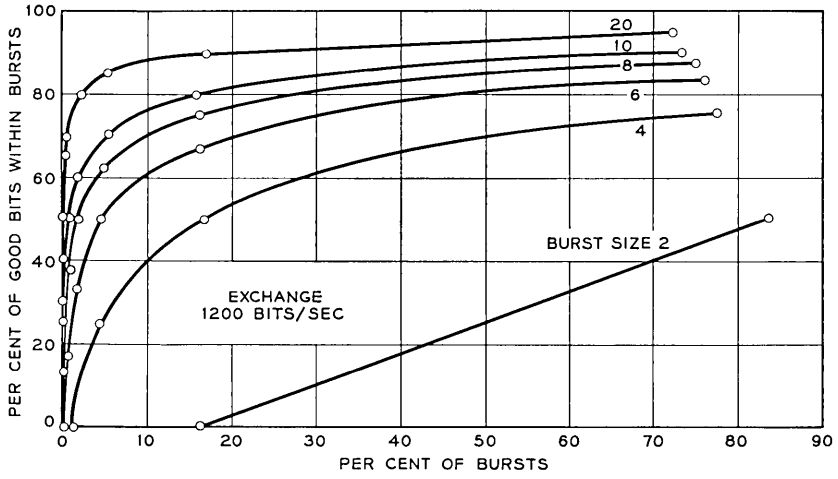


Fig. 39 — Density of good bits within bursts, exchange calls only, 1200 bits per second — percentage of good bits within bursts plotted as a function of the percentage of bursts of various sizes.

of single-error correction codes, to bursts of 20 is shown. This range is provided because it is felt that burst-correcting codes for larger bursts than 20 become quite complex and would be of such high cost that there would be little application for such systems. The curves are shown for all calls and for only the exchange calls, showing how the effectiveness of error-correcting schemes may vary for different types of calls.

To illustrate the improvement that can be expected from a Hagelbarger code, which is designed for correcting bursts of eight bits in duration and which requires a clear-out interval of 26 bits between bursts, an approximation is made. Such a code would have a redundancy of 50 per cent. In Fig. 32, 69 per cent of the bursts for burst length 8 have 26 or more good bits between them. This means these are correctible bursts. If it is assumed that the uncorrected bursts have the same error density as the corrected bursts — namely, that shown by Fig. 33 — then an improvement or reduction in average error rate of about 3.2 to 1 can be expected. The only reason why this is an approximation rather than an exact evaluation is because of the previously stated assumption regarding density of uncorrectible bursts, and also because the coding scheme may introduce additional errors when the bursts are too close. For exchange calls, based on the information shown in Fig. 34, approximately 96.5 per cent of the bursts are correctible by an eight-bit burst-correcting code with 50 per cent redundancy, which should result in an improve-

ment of about 28 to 1. This is because on exchange calls there are fewer uncorrectible bursts, since there are fewer bursts that extend beyond eight bits and fewer bursts that are closer together than 26 bits. It is interesting to note that, if an evaluation is made of a single-error-correcting code that requires, say, 10 good bits between errors (a Hamming code would accomplish this), then it is found that on these exchange calls the single-bit errors predominate. Thus, a substantial amount of the improvement made with an eight-bit burst-correcting code could have been made with a single-error-correcting code.

Now we shall examine all the calls to explore the amount of improvement that may be expected by increasing the error-correcting capabilities from 8 to 20 bits. This means that for the same redundancy the clear-out interval must be extended from 26 to 62 bits. The curves indicate that there is very little additional improvement. Fig. 32 shows that the 20-bit bursts with a clear-out interval of 62 bits represent 79 per cent of the total burst instead of 69 per cent. Therefore, little advantage is obtained compared to the increased circuit complexity that must be provided. These bursts may seem long for data transmission, since they may effect many bits, but for speech the circuits are very satisfactory and such interruptions are rarely noticeable.

Information is provided for determining the effectiveness of these codes for different types of calls. However, the relative value of these coding schemes can better be illustrated on cumulative-error-rate distribution curves similar to those previously described in Figs. 26 and 29. A computer was programmed to correct those errors that were single errors with more than 10 good bits between them, and also was programmed to correct those bursts that did not exceed 8 bits in duration and had at least 26 good bits between them. These values were chosen since they are thought to represent coding systems that can be implemented with relative ease and illustrate order of magnitude improvements that might be expected. The cumulative distributions of uncorrectible errors are shown in Figs. 40 and 41 for speeds of 600 and 1200 bits per second, respectively. Also, plotted on the same axes are the identical curves previously shown in Figs. 26 and 29, which are distributions for these calls without error correction of any type. Thus, it is shown that at 600 bits per second 80 per cent of the circuits achieve an error rate better than one error in more than 100,000 bits, without any error correction. With single-error correction, 85 per cent of the circuits perform this well, and with burst correction the percentage is increased to 91 per cent. It is necessary to keep in mind that, with error correction, redundancy is added and, in the case of burst correction, 50 per cent of

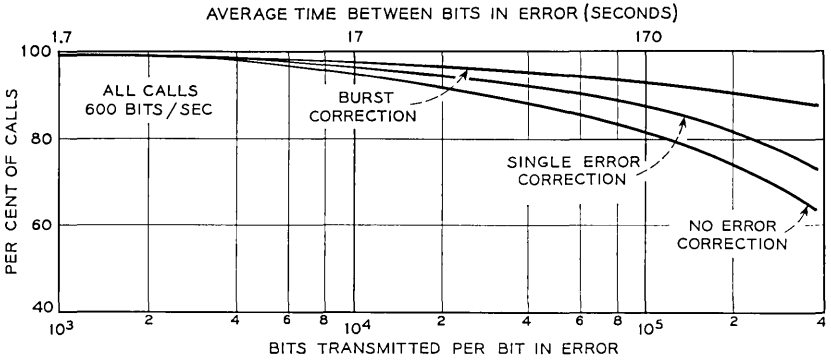


Fig. 40 — Error rate distribution, all calls, 600 bits per second — percentage of calls having an average error rate better than that shown on abscissa.

the bits are check bits. Therefore, a comparison is made of one in 10^5 bits with no correction, to one in 1.7×10^5 bits with single-error correction, and one in 2×10^5 bits with the burst correction.

At 1200 bits per second the improvement in performance with single-error correction and burst correction is somewhat better than at 600 bits per second. Actually, the addition of error control tends to make the performance at 600 and 1200 bits per second very close. For example, at 600 bits per second with burst correction, 94 per cent of the circuits produce an error rate better than one bit per 50,000 transmitted. At

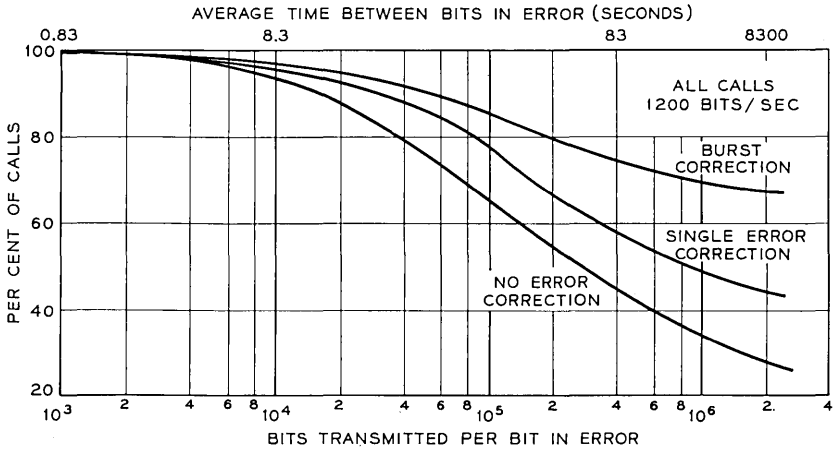


Fig. 41 — Error-rate distribution, all calls, 1200 bits per second — percentage of calls having an average error rate better than that shown on abscissa.

1200 bits per second with the same error control, 90 per cent of the circuits gave the same performance. The sets of curves in Figs. 40 and 41 are for all calls made in the investigation, and therefore include exchange, short-haul and long-haul connections. It is emphasized that if such curves were shown for only exchange calls the improvement would be greater, whereas if they were shown for only long-haul calls the improvement would be less.

These error statistics indicate that, where a high degree of accuracy is required, retransmission of data is also required. Forward-acting error-correcting codes by themselves do not at present appear to be the complete solution. Undoubtedly, progress will be made in the direction of achieving large volumes of storage at low cost, which will facilitate more economical design of forward-acting error-control schemes. Also, as new transmission systems are developed and improvements are made to existing systems, the probability of large bursts of errors will be reduced.

The previous curves have provided the information necessary to aid in making decisions as to whether error control is necessary and what type is most effective. If retransmission is necessary the question then arises as to the optimum block length. An important factor adding to the complexity of this problem is the turnaround of the echo suppressors, the propagation time, and the physical and electrical design of the data input and output machinery. Retransmission methods can cover a vast range of possibilities. For example, one method might be to send blocks of data of just a few bits in duration three times consecutively and take the best two out of three. Another scheme, which might represent the opposite extreme, would be to transmit entire messages, say 10 minutes in duration, and when an error is encountered retransmit the whole message over again. To evaluate the effectiveness of these schemes it is necessary to know the probability of error-free transmission as a function of message length. Also, if this latter scheme were used with single-error correction, so that retransmission would not be required on the single errors but only on the long bursts, this method of error control might be considerably more promising.

Figs. 42 and 43 describe the probability of error-free transmission for no-error correction, single-error correction and eight-bit burst correction. Because of the vast time scales that may be of interest for error-control purposes, the curves are plotted on two different scales. The probability scale on the left permits accurate evaluation for message lengths of less than 1000 bits and the scale on the right is used for longer message lengths. These curves are shown for both 600-bit-per-second and 1200-bit-per-second tests.

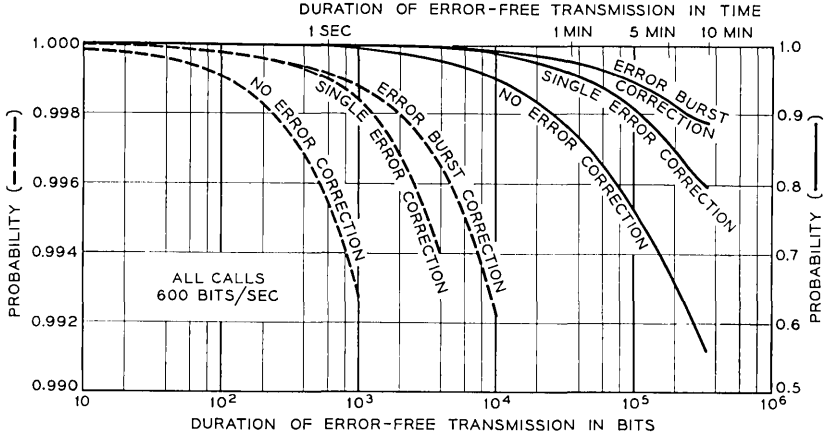


Fig. 42 — Probability of error-free transmission, all calls, 600 bits per second — probability of transmitting as many as or more error-free bits than that shown on abscissa.

For example, Fig. 42 indicates that, with 1000-bit blocks at 600 bits per second with no error correction the probability of error-free transmission is 0.993. With single-error correction this probability increases to 0.9984, and with burst correction this probability further increases to 0.9988. Thus, it is quite obvious that in this application there is very little advantage in forward-acting error-correction codes. A forward-

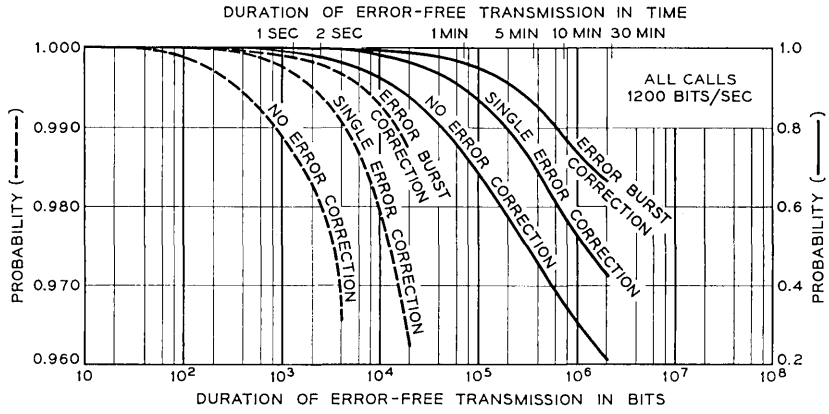


Fig. 43 — Probability of error-free transmission, all calls, 1200 bits per second — probability of transmitting as many as or more error-free bits than that shown on abscissa.

acting error-control scheme may not by itself appear very promising when it provides for a reduction in error rate by a factor, say, of 5 to 1. But if this error-correction scheme is used along with a retransmission method and the forward-acting code reduces the number of retransmissions, then this code may have proved itself by increasing the efficiency in the use of the telephone circuit. Many more interesting examples of error control could be discussed on the basis of these curves, but the objective herein is to illustrate the engineering value of the statistics and let individual ingenuity go to work.

VII. CONCLUSION

The evaluation program has demonstrated that speeds as high as 1200 bits per second with an FM modem using a zero-crossing detection system are entirely practicable on the regular switched telephone network. The error performance on the connections is variable, depending upon a number of factors. In many cases, the probability of error in transmission may be so much lower than the probability of error from other sources that error control may not be necessary. When very high accuracy is required, error-control techniques can be used effectively.

Error-detection and block-retransmission methods appear necessary in order to obtain a high degree of accuracy on long distance transmission. Forward-acting error-correcting codes may be used to improve the line efficiency when such methods are used.

It is possible to design around many of the data limiting characteristics of the network — the compandors and echo suppressors, for example. The variability in circuit characteristics can also be compensated for somewhat by corrective devices associated with either the data terminal equipment or, in some cases, the telephone channel itself. The compromise equalizers used for the 1200-bit-per-second tests are typical examples of what can be done in this direction.

For some applications, arrangements may be made to bypass certain facilities that limit the transmission of data signals. These may take the form of controlled access to the long distance switching network or perhaps the use of only certain telephone facilities and offices in the data service offering. In any case, the final decision as to the engineering design will be determined by the over-all economics.

The Bell System has a continuing effort to achieve higher speeds and greater accuracy, to provide more effective means for handling the variety of data transmission requirements and to broaden the scope of data processing applications by reducing the cost of transmitting information.

VIII. ACKNOWLEDGMENTS

The authors would like to express their appreciation and give special recognition to the following men in Bell Telephone Laboratories who have made major contributions to the success of the measurement program: L. F. Kelley and N. E. Snow, for the design and construction of the specialized test equipment; G. J. McAllister, for magnetic tape processing and computer analysis of error statistics; and L. F. Bugbee and H. J. Ewoldt, for computer analysis of transmission characteristics.

In addition, the enthusiastic cooperation of the many people in the operating companies of the Bell System, the American Telephone and Telegraph Company and Bell Telephone Laboratories is gratefully acknowledged.

REFERENCES

1. Duncan, J. A., Parker, R. D. and Pierce, R. E., Telegraphy in the Bell System, A.I.E.E. Trans., **63,** 1944, p. 1032.
2. Carter, C. W., Jr., Dickieson, A. C. and Mitchell, D., Application of Compandors to Telephone Circuits, A.I.E.E. Trans., **65** (Suppl.), 1946, p. 1079.
3. Clark, A. B. and Mathes, R. C., Echo Suppressors for Long Telephone Circuits, A.I.E.E. Trans., **44,** 1925, p. 481; Horton, A. W., The Occurrence and Effect of Lockout Occasioned by Echo Suppressors, B.S.T.J., **17,** 1938, p. 258.
4. Inglis, A. H. and Tuffnell, W. L., An Improved Telephone Set, B.S.T.J., **30,** 1951, p. 239.
5. Huntley, H. R., The Present and Future of Telephone Transmission, Elec. Engg., **75,** 1956, p. 686.
6. Wilkinson, R. I., Beginnings of Switching Theory in the United States, Elec. Engg., **75,** 1956, p. 796.
7. Pilliod, J. J., Fundamental Plans for Toll Telephone Plant, A.I.E.E. Trans., **71,** Part 1, 1952, p. 248.
8. Clark, A. B. and Osborne, H. S., Automatic Switching for Nationwide Telephone Service, A.I.E.E. Trans., **71,** Part 1, 1952, p. 245.
9. Truitt, C. J., Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks, B.S.T.J., **33,** 1954, p. 277.
10. Newell, N. A. and Weaver, A., Single-Frequency Signaling System for Supervision and Dialing Over Long Distance Telephone Trunks, A.I.E.E. Trans., **70,** Part 1, 1951, p. 489.
11. Newell, N. A. and Weaver, A., In-Band Single-Frequency Signaling, B.S.T.J., **33,** 1954, p. 1309.
12. Bogan, L. B. and Young, K. D., Simplified Transmission Engineering in Exchange Cable Plant Design, A.I.E.E. Trans., **73,** Part 1, 1954, p. 498.
13. Shaw, T., The Evolution of Inductive Loading for Bell System Telephone Facilities, B.S.T.J., **30,** 1951, p. 149.
14. Affel, H. A., Demarest, C. S. and Green, C. W., Carrier Systems on Long Distance Telephone Lines, B.S.T.J., **7,** 1928, p. 564.
15. Fisher, H. J., Almquist, M. L. and Mills, R. H., A New Single-Channel Carrier Telephone System, B.S.T.J., **17,** 1938, p. 162.
16. Starbird, L. C. and Mathis, J. D., Some Applications of the Type J Carrier System, B.S.T.J., **18,** 1939, p. 338.
17. Green, C. W. and Green, E. I., A Carrier Telephone System for Toll Cables, B.S.T.J., **17,** 1938, p. 80.
18. Elmendorf, C. H., Ehrbar, R. D., Klie, R. H. and Grossman, A. J., The L-3 Coaxial System, A.I.E.E. Trans., **72,** Part 1, 1953, p. 395.

19. Caruthers, R. S., The Type N-1 Carrier Telephone System: Objectives and Transmission Features, B.S.T.J., **30,** 1951, p. 1.
20. Edwards, P. G. and Montfort, L. R., The Type O Carrier System, B.S.T.J., **31,** 1952, p. 688.
21. Fracassi, R. D. and Kahl, H., Type ON Carrier Telephone, A.I.E.E. Trans., **72,** Part 1, 1953, p. 713.
22. Roetken, A. A., Smith, K. D. and Friis, R. W., The TD-2 Microwave Radio Relay System, B.S.T.J., **30,** October 1951, p. 1041.
23. Weber, L. A., FM Digital Subset for Data Transmission Over Telephone Lines, Comm. & Electronics, no. 40, 1959, p. 867.
24. Edson, J. O., Flavin, M. A. and Perry, A. D., Synchronous Clocks for Data Transmission, Comm. & Electronics, no. 40, 1959, p. 832.
25. Nyquist, H., Certain Topics in Telegraph Transmission Theory, A.I.E.E. Trans., **47,** 1928, p. 617.
26. Lattimer, I. E., The Use of Telephone Circuits for Picture and Facsimile Service, Long Lines Dept., American Telephone and Telegraph Company, New York, 1948.
27. Mertz, P., Transmission Line Characteristics and Effects on Pulse Transmission, Proc. Symp. on Information Networks, Polytechnic Inst. of Brooklyn, April 1954.
28. Wood, F. B., Optimum Block Length for Data Transmission with Error Checking, Comm. & Electronics, no. 40, 1959, p. 855.
29. Brown, A. B. and Meyers, S. T., Evaluation of Some Error-Correcting Methods Applicable to Digital Data Transmission, I.R.E. Conv. Rec., Part 4, 1958, p. 37.
30. Hamming, R. W., Error Detecting and Error Correcting Codes, B.S.T.J., **29,** 1950, p. 147.
31. Hagelbarger, D. W., Recurrent Codes — Easily Mechanized, Burst-Correcting Binary Codes, B.S.T.J., **38,** July 1959, p. 969.

High-Frequency Negative-Resistance Circuit Principles for Esaki Diode Applications

By M. E. HINES

(Manuscript received January 21, 1960)

Certain fundamental principles are presented for analyzing and designing high-frequency amplifiers and oscillators utilizing simple negative resistance elements such as the Esaki or tunnel diodes. The first part of the paper covers the conditions necessary for oscillation and amplification with a single negative-resistance diode, including stability criteria, gain and bandwidth. It is shown that the highest-frequency circuits require diodes with very small dimensions, so that a single-spot diode will have a very low power capacity. In order to obtain higher power at high frequencies, distributed circuits must be used, either with narrow-strip diodes or a multiplicity of small spot diodes. Such circuits present special stabilization problems in suppressing unwanted modes of oscillation. Methods of avoiding such difficulties are presented for one-port oscillator circuits and for traveling-wave amplifier circuits. In the latter case, nonreciprocal attenuation of the gyromagnetic type is recommended.

I. INTRODUCTION

The Esaki[1] diode (or tunnel diode) exhibits a negative-resistance characteristic in the forward-biased region as shown in Fig. 1. This is a "voltage-controlled" type of characteristic in that the current is a single-valued function of voltage. Fig. 1 is a static curve, but the negative resistance is believed to remain effective at extremely high frequencies. Oscillation in the microwave range has been observed by several workers.[2,3,4] It is suspected that useful negative-resistance effects will also become obtainable in the millimeter wave range as our technology improves.

There is also a capacitance across the junction. This is quite high by comparison with other junction diodes, when measured per unit area

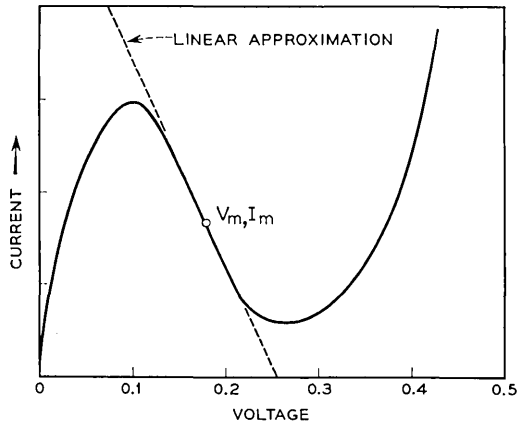


Fig. 1 — The current-voltage characteristic of an Esaki diode. The general curve shape is variable to a small degree, and the current density to a large degree, depending upon the semiconductor materials and processing. The total current is also proportional to diode junction area.

of the junction. The negative conductance, however, is also high, so that the negative time constant (negative R times C) is usually substantially less than 10^{-9} seconds. Negative time constants on the order of 10^{-11} to 10^{-12} seconds are believed to be obtainable in special diodes using intermetallic semiconductor compounds such as indium antimonide.[5]

The basic purposes of this paper are to evaluate the Esaki diode principle as a useful element in practical microwave devices, to show its limitations and capabilities and to present certain elementary device design methods for microwave oscillators and amplifiers. We will discuss only the circuit aspects of Esaki diodes as negative-resistance devices at high frequency. The solid state physics of the tunneling process applicable to these diodes has been described by others.[1,6,7] Microwave devices must include substantial parts of the high-frequency circuits, in a manner similar to that of microwave electron tubes. Suitable circuit geometries will be described and analyzed, taking into account the junction capacitance, negative resistance, parasitic resistance, load coupling, etc. This work is mostly theoretical and little experimental work is described.

The paper begins with a discussion of the stability criteria for simple negative-resistance circuits, which presents the appropriate relationships among the circuit parameters for amplification, oscillation and switching. The analysis includes the limiting effects of circuit and load resistance, diode capacitance and negative resistance. It is shown that

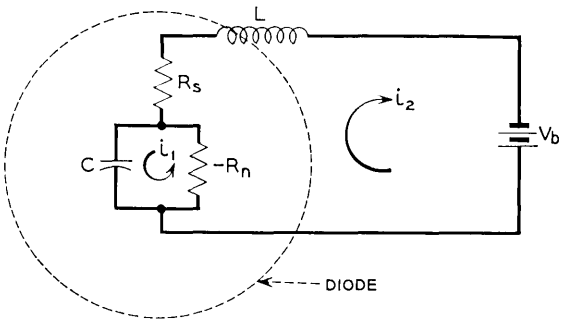


Fig. 2 — A simple equivalent circuit of an Esaki diode connected to a battery. Analysis for stability must include the lead inductance L, the parasitic resistance R_s and the capacitance C, as well as the negative resistance R_n .

very small diode dimensions will be required for high microwave frequencies.

If appreciable microwave power is to be obtained, distributed circuits will be required. These can take the form of extended narrow-strip diode junctions or a multiplicity of small-spot diodes in a filter-type structure. Such circuits pose difficult problems in device design and fabrication, and special precautions are necessary to avoid oscillation in spurious resonant modes. A substantial part of this paper is devoted to the latter problem, and several circuit possibilities are described.

II. SIMPLE NEGATIVE-RESISTANCE CIRCUITS

2.1 *Basic Stability Criteria*

We will begin with the simplest possible circuit, which is simply a diode connected to the terminals of a battery. This circuit, shown in Fig. 2, also may be interpreted to include a number of practical circuits in which additional inductance and resistance have been added. A meaningful analysis must include the finite lead inductance, the inductance of the battery loop, and at least the inherent passive resistance of the battery and diode.

R. L. Wallace has derived (4) below in unpublished work. We will repeat this and give a concise interpretation in the form of a stability diagram.

The V–I curve of Fig. 1 is nonlinear. We use a linearizing approximation valid in the immediate vicinity of the operating point V_m , I_m , shown in Fig. 1,

$$i = I_m - \frac{V - V_m}{R_n}, \tag{1}$$

where i and V are the instantaneous current and voltage, and the negative resistance R_n is the inverse of the V–I slope at the operating point. In a straightforward manner, one can write two loop equations and make appropriate elementary substitutions to obtain a differential equation for the current i_2 in the battery loop. This is

$$-L\frac{d^2i_2}{dt^2} + \left(\frac{L}{R_nC} - R_s\right)\frac{di_2}{dt} + \frac{R_s - R_n}{R_nC}\,i_2 = \frac{1}{R_nC}\,(V_m - V_b + R_nI_n). \tag{2}$$

The general solution of the above equation is

$$i_2 = A_1e^{p_1t} + A_2e^{p_2t} + \frac{V_b - V_m - R_nI_m}{R_s - R_n}, \tag{3}$$

where the third term is the dc bias current and p_1 and p_2 are the two values (taking the + and − signs) given below.

$$p_{1,2} = \frac{1}{2}\left(\frac{1}{R_nC} - \frac{R_s}{L}\right) \pm j\sqrt{\frac{1}{LC}\left(1 - \frac{R_s}{R_n}\right) - \frac{1}{4}\left(\frac{R_s}{L} - \frac{1}{R_nC}\right)^2}. \tag{4}$$

In (3), A_1 and A_2 are arbitrary constants depending upon the initial current in the inductor and charge on the capacitor. The exponential factors p_1 and p_2 may be real, complex or imaginary, depending upon the choice of circuit parameters. If either value has a positive real part, the circuit will be unstable. If the p's are real, an initial disturbance will either grow or decay exponentially to the steady bias condition. If the p's are complex, the transient waves will be growing or decaying sinusoids.

Equation (4) has four parameters. We can reduce this to a two-parameter function by the substitutions

$$\omega_0 = \frac{1}{\sqrt{LC}} \tag{5}$$

and

$$Q_n = \omega_0R_nC = \frac{R_n}{\sqrt{\frac{L}{C}}}. \tag{6}$$

These yield

$$\frac{p}{\omega_0} = \frac{1}{2Q_n}\left(1 - \frac{R_s}{R_n}\,Q_n^2\right) \pm j\sqrt{\left(1 - \frac{R_s}{R_n}\right) - \frac{1}{4Q_n^2}\left(1 - \frac{R_s}{R_n}\,Q_n^2\right)^2}. \tag{7}$$

Fig. 3 shows a set of curves for the above function, and Fig. 4 is a stability diagram suggested by W. W. Anderson. These show that the circuit will be unstable if the ratio of R_s to R_n is either too small or too large, or if the ratio of inductance to capacitance is too large. For many Esaki diodes, the negative conductance is so large that the inductance of the shortest pigtail leads is sufficient to cause oscillations, and a special low-inductance case is necessary to allow stable biasing at the operating point. For operation at VHF or lower frequencies, it is usually necessary to add capacitance in order to permit a lower inductance, and thereby get a value for Q_n that is large enough to obtain sinusoidal operation as opposed to exponential or "blocking" type oscillations. A typical value for the product R_nC might be 2×10^{-10} in a germanium Esaki diode. For such a diode the maximum allowable inductance for sinusoidal oscillations would resonate with the capacitance at about 400 mc. If a

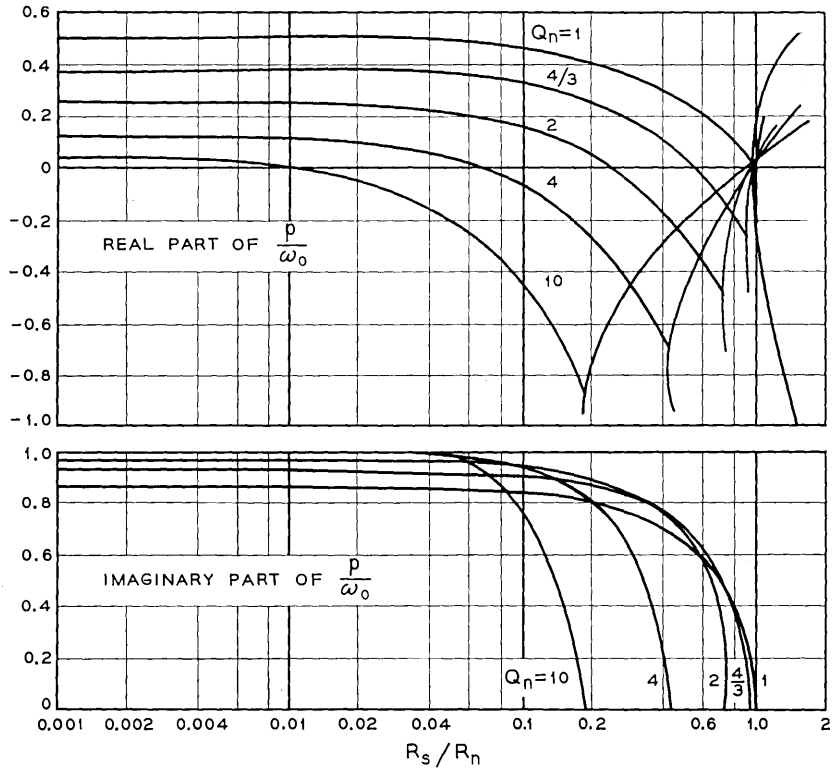


Fig. 3 — Exponential transient characteristics for the circuit of Fig. 2. Transient currents vary as e^{pt}. Here ω_0 is the resonant frequency of L and C, and Q_n is defined as $\omega_0 R_n C$.

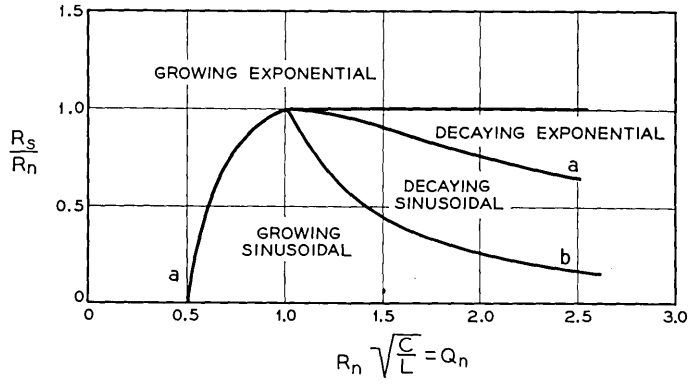


Fig. 4 — A regional stability chart for the circuit of Fig. 2, showing the allowed ranges of the parameters for particular types of transient waves. For curve a, $R_s/R_n = 2/Q_n - 1/Q_n^2$; for curve b, $R_s/R_n = 1/Q_n^2$.

stable condition is required, the resonant frequency must be at least twice this value, and preferably several times higher. If extra capacitance is to be added, it must be connected without adding appreciable inductance between the diode and the capacitor. One method would be to include the capacitance in the diode case or to mount the diode between the plates of a capacitor.

2.2 *An RF Circuit with External Battery*

It is seldom possible or desirable to include the battery or power supply in the high-frequency portion of an RF circuit. It is possible to isolate the RF region through the use of a bypass capacitor and an RF choke, but special precautions are necessary to avoid instabilities because of the inductance of the choke or of power leads alone. The use of capacitance to stabilize negative-resistance circuits has been described by Thomas.[8]

Fig. 5 shows a simple and useful RF circuit with such isolation; the parts drawn with heavier lines are the RF region, with C_2 being a large bypass capacitor and L_2 the inductance of the power leads and any added choke coil. The load resistance R_L should be inductively coupled, either as shown or by mutual inductance through a transformer. A stabilizing resistance, R_{s2}, may simply be the internal impedance of the battery or power supply. The ranges of values for L_2, C_2 and R_{s2} are strictly limited; otherwise we may expect instabilities at low frequencies.

To analyze the circuit of Fig. 5, we assume that C_2 is very large and

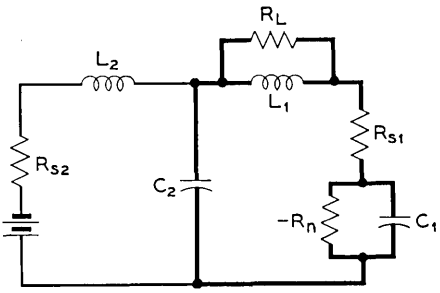


Fig. 5 — An RF Esaki diode circuit in which the RF and power supply are isolated at high frequencies. The bypass capacitor C_2 serves to confine the RF currents to the part of the circuit drawn with heavier lines. It is also helpful in stabilizing the circuit against instabilities involving the inductance of the power leads. Low-frequency stabilization can be further helped by shunting a low resistance across the capacitor C_2.

break the circuit into separate high-frequency and low-frequency equivalents as seen from the RF and dc terminals respectively. These are shown in Fig. 6. The circuits are in the same form as in Fig. 2, and we may use the criteria of Fig. 4 and (7) to determine the stability of each circuit and the nature of the transient waves involved. In the RF equivalent circuit we have replaced the load resistance by its series equivalent and assumed that C_2 is an RF short circuit. For the low-frequency equivalent circuit, we have assumed that L_1 is a short circuit. Four transient wave types will be obtained from these two circuits, and four would be obtained by an exact analysis of the complete circuit. We can presume that the results from the two equivalent circuits are close approxima-

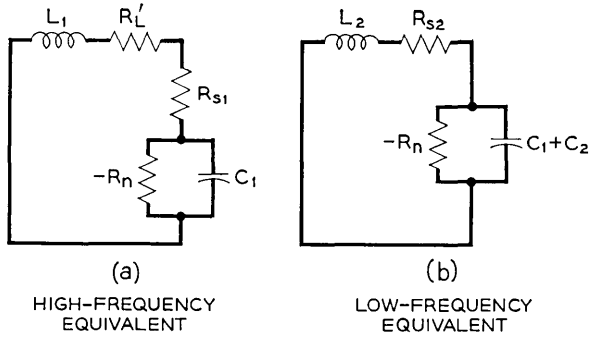


Fig. 6 — Two equivalent circuits applicable to Fig. 5. These may be separately analyzed by the criteria of Fig. 4 to determine the four transient wave types for the circuit of Fig. 5.

tions, provided that the values of p obtained are consistent with the assumptions about the low- and high-frequency impedance of L_1 and C_2 that allowed us to separate the circuits. For the low-frequency equivalent, the criteria for stability are

$$R_n \sqrt{\frac{C_1 + C_2}{L_2}} > 1, \tag{8}$$

$$R_n > R_{s2} > \frac{L_2}{C_1 + C_2} \frac{1}{R_n}. \tag{9}$$

It is possible to use a small stabilizing resistance R_{s2} provided that C_2 can be sufficiently large compared to L_2.

Fig. 7 shows a coaxial cavity version of the circuit we have been discussing in this section. An actual circuit of this type has been used by A. Yariv and E. Dickten at Bell Telephone Laboratories to obtain oscillation at frequencies above 8000 mc, using a germanium Esaki diode.

2.3 *Oscillation Conditions*

The conditions for obtaining oscillations can be deduced from the stability criteria; this requires a transient wave with a positive real part for p. If p is also complex, a growing sinusoidal transient is indicated.

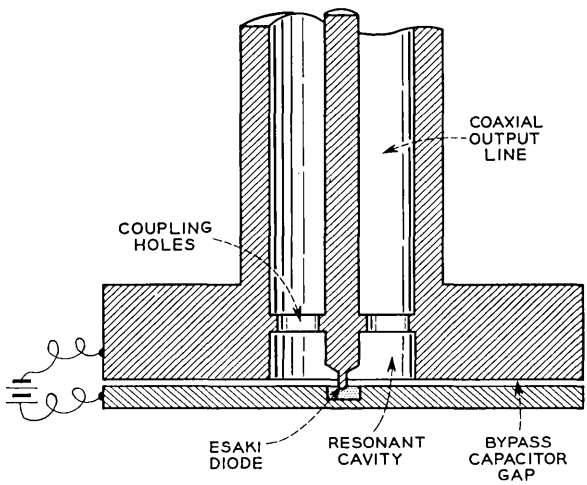


Fig. 7 — A schematic sketch of a coaxial cavity microwave circuit utilizing the isolation principle of Fig. 5. Oscillations at frequencies above 8000 mc have been obtained by A. Yariv and E. Dickten in a circuit of this type using a germanium Esaki diode.

Nonlinearities must limit the maximum growth of such a transient because large voltage swings will extend into the positive resistance region of the diode. Under these conditions, we may usually expect a steady oscillation. The region of growing sinusoidal transients is clearly shown in Fig. 4. Useful oscillations are also obtainable for low values of R_s/R_n and low Q in the lower left part of Fig. 4, where growing exponential transients are indicated. Because of the nonlinearity, we can expect the voltage to "switch" back and forth at high speed. Steady oscillations will not be obtained, however, for $R_s > R_n$. In that case, the circuit tends to stabilize at a voltage either above or below the region of maximum negative conductance. This condition is useful for logic circuitry.

2.4 *Amplification*

For amplification, a number of circuit configurations are possible. Fig. 8 shows the circulator method of obtaining useful gain with the negative resistance circuit of Fig. 5 or Fig. 7. The input wave passes through the circulator to the amplifier, and the reflected wave is diverted by the circulator to pass out by another port. When the impedances are properly adjusted, the circuit will be stable, but the reflected wave will be greater than the incident wave, resulting in a net power gain. Another method of obtaining amplification has been described and analyzed by Chang.[9] His scheme uses input and output lines separately coupled to the negative-resistance circuit.

A third method is to use a 3-db directional coupler or a hybrid junction (magic tee) with two negative-resistance circuits. This method is shown in Fig. 9. The waveguides connecting the tee to the amplifiers are unequal in length, with an additional one-fourth wavelength in one arm. This causes the reflections from the two amplifiers to be out of phase at the tee and pass out by the fourth port. Of the three methods, the circulator approach gives unilateral gain, and multiple stages can be used without increasing the danger of oscillation. The hybrid and two-port amplifiers are more sensitive to mismatches at input and output, and isolators are essential if a high total gain is obtained.

In order to analyze the reflection-type amplifier, we may conveniently use a shunt equivalent form, as shown in Fig. 8. Here we have a parallel-resonant combination and have replaced $-R_n$ by its inverse $-G_n$. We also include a shunt positive conductance G_p to account for parasitic losses in the diode and circuit.

The power gain of the circuit is given by the square of the magnitude of the voltage reflection coefficient Γ. This is the familiar expression

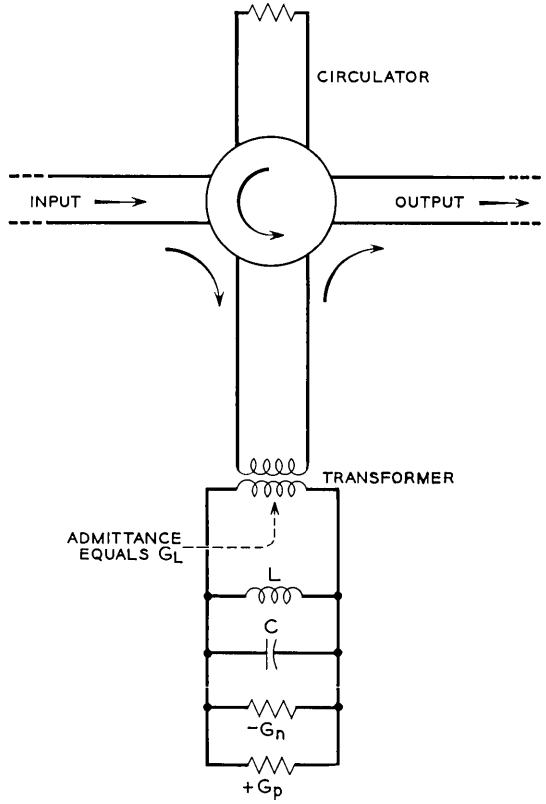


Fig. 8 — The circulator method of connecting a negative-resistance circuit to give useful amplification. The circuit of Fig. 7 is suitable for this kind of amplifier. An equivalent circuit in shunt form is shown for the amplifier in this figure.

$$\Gamma = \frac{G_L - Y}{G_L + Y}, \tag{10}$$

where G_L is the characteristic admittance of the connecting transmission line and Y is the admittance of the "amplifier" circuit that forms the terminating admittance of the line. For the circuit of Fig. 8,

$$Y = G_p - G_n + j\omega C - \frac{j}{\omega L}, \tag{11}$$

so that the power gain g_p will be

$$g_p = \left| \frac{G_L - G_p + G_n - j\omega C + \frac{j}{\omega L}}{G_L - G_n + G_p + j\omega C - \frac{j}{\omega L}} \right|^2. \tag{12}$$

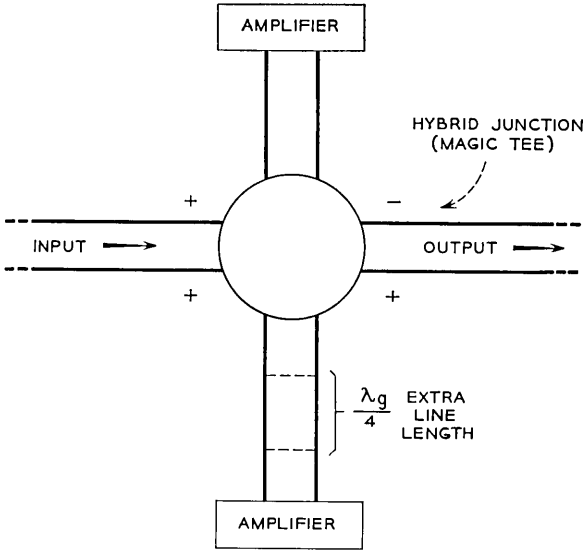


Fig. 9 — The hybrid-junction (magic-tee) method of obtaining amplification using two negative-resistance circuits. The extra length in the lower arm reverses the phase of one reflected signal so that the two reflected waves will combine in the output line of the hybrid junction. A 3-db directional coupler may be used as an alternative to the hybrid, but the proper phase conditions will depend upon the particular coupler used.

The gain will be greater than unity if G_n is greater than G_p and will approach infinity at resonance if G_n approaches G_p plus G_L. Oscillations will occur if G_n is greater than G_p plus G_L.

As in other negative-resistance amplifiers, the bandwidth decreases as the gain is increased. If a simple resonant circuit is used and the gain is high, the product of voltage gain and bandwidth is more or less invariant as the gain is changed over a wide range. At high gain, G_L is approximately equal to $G_n - G_p$, so that at resonance the voltage gain is approximately

$$g_v = \frac{2(G_n - G_P)}{G_L - (G_n - G_P)} \tag{13}$$

and the 3-db bandwidth is

$$B_{3db} = \frac{f_0}{Q_{net}} = f_0 \frac{G_2 - (G_n - G_P)}{\omega_0 C}. \tag{14}$$

The gain-bandwidth product, therefore, is

$$g_v B_{3db} = \frac{G_n - G_P}{\pi C}. \tag{15}$$

In the special case of negligible parasitic resistance,

$$g_v B_{3db} = \frac{G_n}{\pi C} = \frac{1}{\pi R_n C}. \tag{16}$$

As shown by Seidel and Herrmann,[10] however, this limitation does not apply for circuits of greater complexity.

2.5 *Maximum Frequency and Diode Geometry*

At present there are no indications of frequency limitations in the microwave range in the negative resistance of the diode junction itself. We assume here that the only significant frequency limitations are those resulting from diode capacitance and parasitic resistance. These are quite sufficient. In this section we shall derive an expression for the maximum frequency and show how it depends upon the characteristics of the diode junction and upon the diode geometry.

Let us assume that we are successful in obtaining a very small value for the series parasitic resistance R_s and that we are attempting to work at a high frequency where Q_n is also large. In this case, we can assume that

$$R_n \gg \frac{1}{\omega_0 C} \gg R_s. \tag{17}$$

For this condition, we can redraw the circuit as a shunt combination of a capacitor, a positive resistor and a negative resistor. To a close approximation, $-R_n$ and C will remain the same and we can take account of R_s by adding a shunt resistor of value $(\omega_0^2 C^2 R_s)^{-1}$. Thus, $\omega_0^2 C^2 R_s$ is the value of G_p in the circuit of Fig. 8. There is a value of ω_0 for which G_p is equal to G_n or R_n^{-1}, and at higher frequencies G_p will be greater. This is the maximum frequency for negative resistance effects, given by

$$f_{\max} = \frac{\sqrt{\frac{R_n}{R_s}}}{2\pi R_n C}. \tag{18}$$

The denominator of this expression is dependent upon the properties of the junction and is independent of geometry. The numerator can be affected by the mechanical and electrical design of the diode mount and

by the size and shape of the junction. In small alloy-junction diodes (most Esaki diodes have been of this type) most of the parasitic resistance will be found in the semiconductor material in the vicinity of the junction. This is the familiar "spreading-resistance" of point-contact diode theory.

In order to compare the useful types of diode geometry, we assume that this spreading-resistance is the only significant contributor. Actually, skin-effect is also important in many cases, but it can be treated by standard methods. Fig. 10 shows two widely divergent types of diode construction, idealized in shape in order to permit a simple analysis. For these cases we assume that the radius r_s is smaller than the skin depth in the semiconductor material, and that the metallic parts have negligible resistance. The junction has a capacitance of C_{d1} farads per square meter and a negative conductance of $-G_{d1}$ mhos per square meter. The semiconductor material has a resistivity of ρ ohm-meters. From these parameters, it is a straightforward matter to derive expressions for the maximum frequency for the two geometries of Fig. 10.

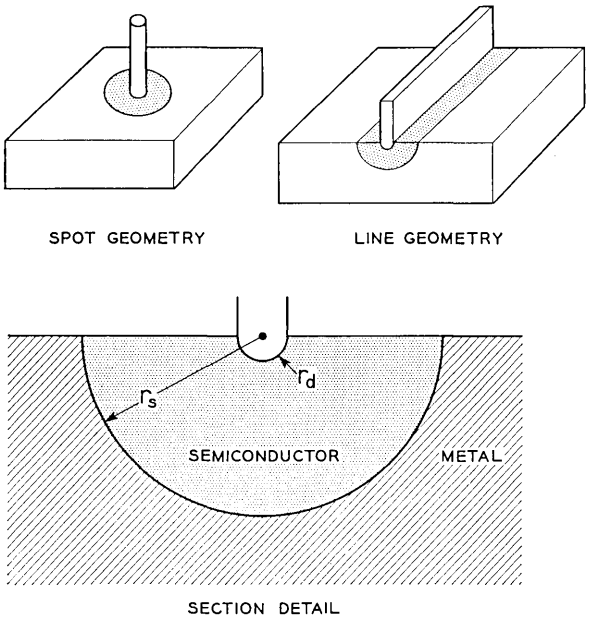


Fig. 10 — Two idealized types of Esaki diode geometry suitable for microwave applications. The circular cross section of the semiconductor is not likely to be a practical structure, but it has been assumed here to simplify the analysis and allow direct comparisons.

For the spot,

$$f_{\max} = \frac{G_{d1}^{\frac{1}{2}}}{2\pi C_{d1}\rho^{\frac{1}{2}}}\left[\frac{1}{r_d\left(1 - \frac{r_d}{r_s}\right)}\right]^{\frac{1}{2}}, \tag{19}$$

and for the strip,

$$f_{\max} = \frac{G_{d1}^{\frac{1}{2}}}{2\pi C_{d1}\rho^{\frac{1}{2}}}\left(\frac{1}{r_d \ln \frac{r_s}{r_d}}\right)^{\frac{1}{2}}. \tag{20}$$

The first term in each expression is the same, and is dependent upon the semiconductor and junction properties only. The second term depends upon geometry only. It is clear that r_d must be small for both cases if the semiconductor is of appreciable thickness. If $r_d/r_s \rightarrow 1$, the frequency limits are equal, but for small r_d/r_s the spot geometry has a higher limit. For example, if $r_d/r_s = 0.1$, the frequency limit for the spot case will be 1.6 times that for the strip case.

III. DISTRIBUTED CIRCUITS — GENERAL CONSIDERATIONS

In Section II the conditions for obtaining useful negative-resistance effects were presented. It was shown that single-spot diodes must be of small area if the highest frequencies are to be usable. It was also shown that narrow-strip diodes can be utilized at high frequencies and that these can have substantially greater power capacity. If significant amounts of power are to be obtained, it will be necessary to use a large number of single-spot diodes or to use narrow-strip diodes of appreciable total length. The remainder of this paper will be devoted to such distributed circuits.

A basic characteristic of distributed circuits in this sense is that more than one resonant mode is possible in the general frequency range of interest. With negative-resistance elements distributed along such a structure, spurious oscillations are a major hazard. This is the familiar *moding* problem which has plagued the development of multicavity magnetrons since the early days of World War II. If we attempt to design ordinary Esaki-diode distributed circuits of equivalent complexity, the problems are likely to be even more serious, because of the broadband nature of the Esaki-type negative resistance. There appear to be two methods of avoiding such difficulties. One is to use circuits of essential simplicity with few diodes, so that the modes are clearly distinct in character and/or frequency and can be separately damped by resist-

ance. The other is to use traveling-wave circuits combined with a nonreciprocal (gyrator) type of attenuation. These are the methods to be discussed here.

IV. TWO-DIODE CIRCUITS

Fig. 11 shows a two-diode circuit for push-pull operation. This circuit has two resonant modes, which we will call the push-pull mode and the unison mode. In the former, the ac voltages are out of phase at the diodes and an ac voltage null is found at the center between the two inductors. The resistance R_s does not affect the Q of this resonance, but the load resistance R_L is energized instead. In the unison mode, the load resistance is not effective, but the alternating currents combine and pass through the resistor R_s. Thus, we have a kind of orthogonality distinguishing the two modes so that they are loaded separately. Fig. 11 shows two

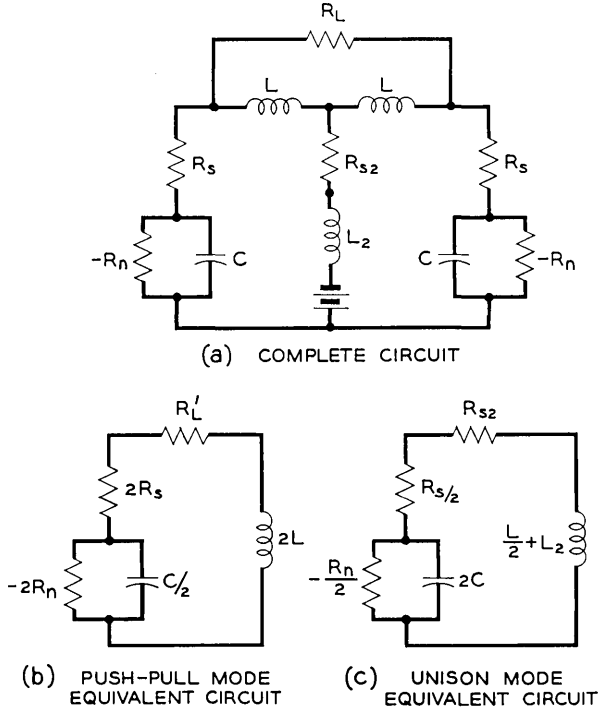


Fig. 11 — A simple push-pull circuit using two Esaki diodes. Two resonant modes are possible, and it is necessary to stabilize the unison mode that involves the power connections. The stability conditions may be determined from the two equivalent circuits shown.

equivalent circuits for the two modes. In Fig. 11(c) the load resistance R_L has been replaced by its series equivalent R_L'. As before, these equivalents can be analyzed for stability by the diagram of Fig. 4. The unison mode must be stable, and the push-pull mode should give a growing sinusoidal transient if oscillations are desired.

Figs. 12 and 13 show two possible microwave applications of the orthogonal mode-separation principle. In Fig. 12 we have two diodes in a double-ended coaxial cavity operating in a "half-wave-length" mode. This circuit is similar to one suggested by R. L. Wallace. The desired push-pull resonance is similar to that of a half-wavelength coaxial line with the ends open-circuited. The actual cavity would be much shorter than a half wavelength because of the excess capacitance at the ends that would be provided by the diodes. Power is fed through a resistor to the middle of the center conductor. This point is a voltage null for the desired push-pull mode. The inductance of this resistor must be included in the equivalent circuit [Fig. 11(c)], and we can analyze the unison mode including the bypass capacitor and power leads in the manner of Fig. 6. Coupling to a waveguide can be accomplished through a window as shown.

Fig. 13 shows a strip-line type of circuit using two diodes, which are mounted on short posts between the plates of the line. The strip line should be sufficiently narrow that it will propagate only the TEM mode at the frequency of interest. For the desired push-pull mode of operation we will have a local resonant circuit involving the inductance of the two posts and the capacitance of the diodes. This mode, if symmetrical, cannot excite a wave in the strip line. The unison mode, however, is

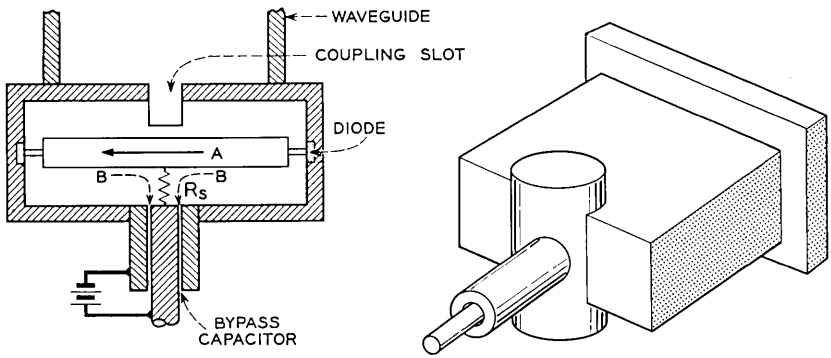
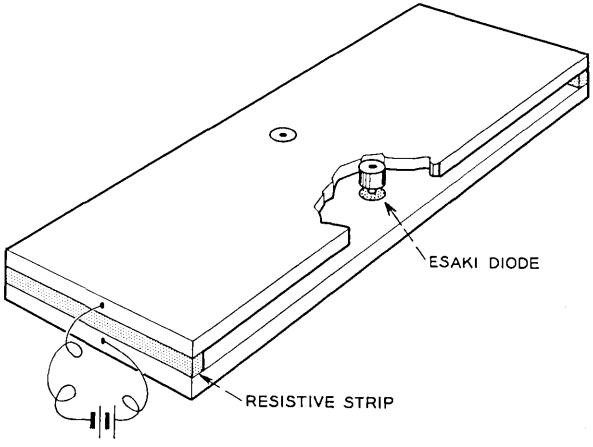
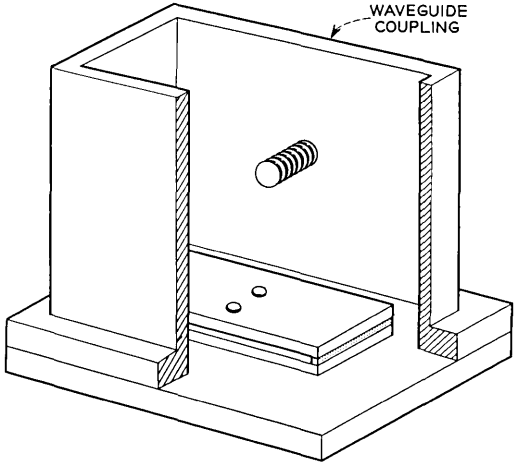


Fig. 12 — A possible coaxial cavity arrangement for a push-pull microwave circuit. The window between the cavity and waveguide provides the loading resistance. The resistance in the cavity stabilizes the unison mode without loading the push-pull mode.

directly coupled to waves on the line, and oscillations can be suppressed by choosing an appropriate matching resistance at a suitable distance from the diode. Also shown, in Fig. 13(b), is a method of mounting the strip line across the end of a waveguide to provide a method of output coupling. If the strip lines are terminated in resistive films at the corners



(a) STRIP-LINE CIRCUIT



(b) METHOD OF MOUNTING

Fig. 13 — A possible push-pull circuit in strip-line form. The strip resistances stabilize the unison mode, but the balanced push-pull mode cannot induce waves along the strip line. The push-pull mode can be coupled to a waveguide if the strip line is placed across the end as shown.

of the waveguide, these resistive films will not be coupled to the waveguide fields. The degree of coupling of the desired mode to the waveguide can be decreased or increased by using a wide or narrow strip line respectively. An alternative method would be to place an iris or post in the waveguide at some distance from the end, as shown.

V. TRAVELING-WAVE DISTRIBUTED CIRCUITS

Fig. 14 shows the simplest possible form of "smooth" distributed circuit. As drawn here, it would be unsuitable for practical use, since it does not include a method of applying dc power and avoiding instability in a uniform-phase mode. This is simply a block of semiconductor with an Esaki-type p-n junction separating p and n regions in the block. The barrier layer between the p and n sides can act as a gap, forming a kind of strip-line.

R. L. Wallace has analyzed this case in an unpublished work, and we will follow his line of attack, using MKS units. We will assume that the junction will act as a parallel-plate transmission line with a very narrow gap. This gap will be the barrier region of the diode, which may be as narrow as 50–100 angstroms. We will simply assume that the gap has a capacitance of C_{d1} farads per square meter and a negative conductance across the gap of $-G_{d1}$ mhos per square meter. Outside the gap on either side we will have a region of poor positive conductance. Skin effect in these regions will have a dominant effect on the transmission characteristics of the line.

As shown in Fig. 14, we assume a strip of width w, and a resistivity of ρ_1 and ρ_2 for the two outer regions. We will be concerned with the propagation of the TEM mode along the strip.

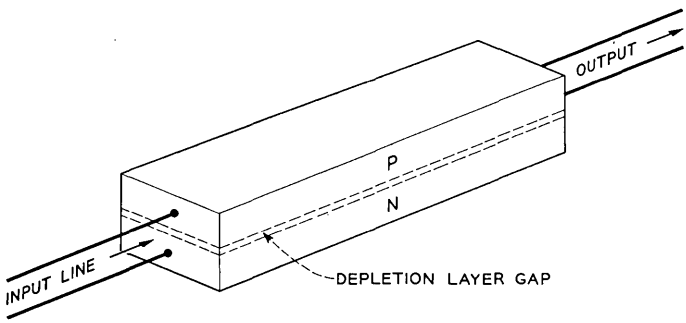


Fig. 14 — An Esaki p-n junction "strip line" that has been analyzed for traveling-wave type gain. This does not appear to be a practical geometry, but the results of the analysis are interesting in considering the stability of large-area Esaki diode junctions.

The familiar formula for skin-effect resistance is

$$R_{sk} = \sqrt{\rho \pi f \mu} \qquad \text{ohms per square,} \tag{21}$$

where ρ is in ohm-meters and μ is the free-space permeability (value: $4\pi \times 10^{-7}$ henrys per meter). There is also an inductive reactance associated with skin effect, which, in planar geometry, has the same magnitude as the resistance:

$$jX_s = j\sqrt{\rho \pi f \mu} \qquad \text{ohms per square.} \tag{22}$$

The inductance associated with the barrier-region gap itself can be neglected, because this gap is very narrow compared to the skin depth. An expression for the series impedance of the line is, therefore,

$$Z = \frac{1}{w}(1 + j)\sqrt{\pi f \mu}(\sqrt{\rho_1} + \sqrt{\rho_2}) \qquad \text{ohms per meter,} \tag{23}$$

where ρ_1 and ρ_2 are the semiconductor resistivities on either side of the junction.

The shunt admittance of the gap as a transmission line is the sum of the capacitive susceptance of the gap plus the negative conductance of the region as an Esaki diode,

$$Y = (jw\omega C_{d1} - G_{d1}w) \qquad \text{mhos per meter,} \tag{24}$$

where C_{d1} and G_{d1} are the capacitance and negative conductance per square meter respectively.

The propagation of waves along a transmission line is given by the expression

$$V(z,t) = V_0 e^{j\omega t \pm \sqrt{ZY}z}, \tag{25}$$

and we are particularly interested in the propagation constant, $\sqrt{ZY}$. If its real and imaginary parts have opposite signs, we can expect gain; if they are of the same sign, we will have loss. From (23) and (24) we obtain

$$\begin{aligned}\sqrt{ZY} &= [(1 + j)\sqrt{\pi f \mu}(\sqrt{\rho_1} + \sqrt{\rho_2})(j\omega C_{d1} - G_{d1})]^{\frac{1}{2}} \\ &= j[\sqrt{\pi f \mu}(\sqrt{\rho_1} + \sqrt{\rho_2})G_{d1}]^{\frac{1}{2}} \\ &\qquad \cdot \left[1 + \frac{\omega C_{d1}}{G_{d1}} + j\left(1 - \frac{\omega C_{d1}}{G_{d1}}\right)\right]^{\frac{1}{2}}.\end{aligned} \tag{26}$$

We will obtain gain if the real and imaginary parts of $\sqrt{ZY}$ are of opposite sign, which requires the imaginary part inside the second brackets to be positive. Thus, the highest frequency for gain is

$$f_{max} = \frac{G_{d1}}{2\pi C_{d1}}. \tag{27}$$

This frequency is independent of the strip width (for our assumed wide strip) and depends only upon the properties of the junction and the semiconductor materials involved. In comparison with the results of the spot and narrow-line geometry cases, we are much more severely limited here.

It is a straightforward matter to compute the wavelength, gain per wavelength and "characteristic impedance" of this kind of structure as a transmission line. These are given by the expressions

$$\text{wavelength} = \frac{2\pi}{\text{imaginary part of } \sqrt{\overline{ZY}}} \text{ meters}, \quad (28)$$

$$\text{gain per wavelength} = 54.7 \frac{\text{real part of } \sqrt{\overline{ZY}}}{\text{imaginary part of } \sqrt{\overline{ZY}}} \text{ db}, \quad (29)$$

$$\text{characteristic impedance} = \sqrt{\frac{\overline{Z}}{\overline{Y}}} \text{ ohms (will be complex)}. \quad (30)$$

A little computation will show that the wavelength will be much shorter than that in free space, that the gain will be very high (for frequencies below the maximum frequency) and that the characteristic impedance will be very low if the strip has appreciable width.

This wide-strip diode geometry does not appear to be particularly interesting at this time for a useful traveling-wave device. The frequency limitation is relatively low, the impedance will be *very* low and there will be serious stability problems in providing a suitable power feed.

A very-narrow-strip diode geometry appears to be more promising. One method of using such a strip diode would be as shown in Fig. 15. Here we have a long, narrow negative-resistance diode placed along the center line of a metallic strip line of appreciable width. At high frequencies, this will propagate a growing wave whose phase velocity will be slower than the velocity of light because of the capacitance loading of the diode. For such slow wave propagation, the magnetic and electric fields in the gap will decay exponentially in the transverse direction away from the diode strip. The field patterns are shown in Fig. 16. At high frequencies, the fields will decay more rapidly with transverse distance than at low frequencies. Thus, a resistive shunt can be placed continuously along both outer edges of the strip line where the high-frequency fields will be weak, and this will not cause a serious amount of high-frequency attenuation. At low frequencies, however, this shunt can cause very substantial attenuation, and if the net positive conductance per unit length exceeds the net negative conductance there will be no problem with low-frequency stability or in connecting the power leads.

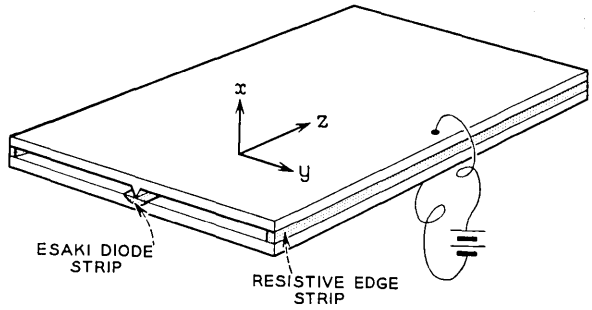


Fig. 15 — A possible traveling-wave Esaki diode amplifier using a narrow-line junction diode. The edge resistances can stabilize the circuit at low frequencies and allow the connection of power leads. At high frequencies, the ac fields are confined to the region near the diode strip, so that transmission is little affected by the resistance strips.

This mode of propagation will be substantially different from the TEM mode of a simple strip line with a central *metallic* ridge with a narrow gap, for which we would expect no velocity reduction. This diode strip has a capacitance corresponding to the barrier gap that is in the order of 10^{-6} cm, while the inductance for currents along the diode

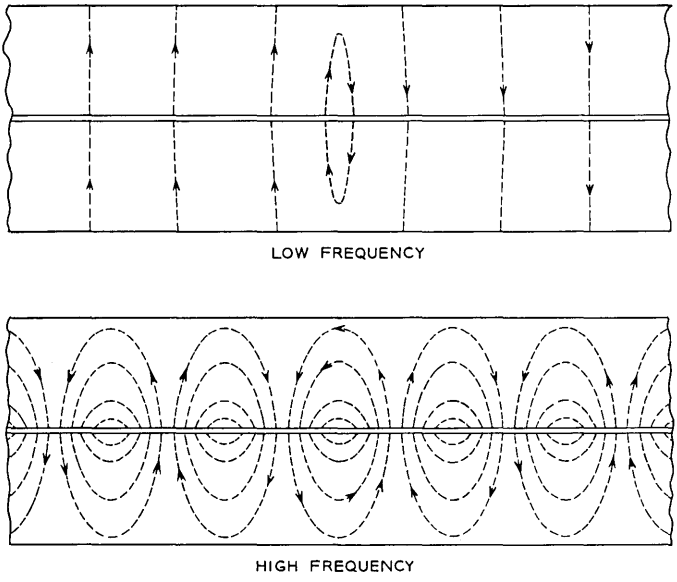


Fig. 16 — The ac magnetic field configuration in the gap region of the circuit of Fig. 15. At high frequencies the ac fields decay exponentially with distance from the center line.

strip corresponds to a gap equal to the skin depth in the semiconductor which is in the order of 10^{-2} to 10^{-3} cm. Thus, propagation would more closely resemble that for a strip line loaded capacitively with a central web of dielectric with a high dielectric constant.

The field pattern in the gap region outside the diode will be a TE mode. If we assume perfectly conducting walls, the field equations in MKS units are

$$H_z = Ae^{j\omega t-\gamma z+qy} + Be^{j\omega t-\gamma z-qy}, \tag{31}$$

$$H_y = \frac{\gamma q}{\gamma^2 + k^2}(Ae_{as}^{j\omega t-\gamma z+qy} - Be^{j\omega t-\gamma z-qy}), \tag{32}$$

$$E_x = \frac{j\omega\mu q}{\gamma^2 + k^2}(Ae^{j\omega t-\gamma z+qy} - Be^{j\omega t-\gamma z-qy}), \tag{33}$$

where z is the direction of propagation, x is perpendicular to the plane of the strips and y is transverse in the plane of the strips. In the above, γ is the complex propagation constant we wish to determine, and

$$k^2 = \omega^2\mu\epsilon, \tag{34}$$

$$q = \pm j\sqrt{\gamma^2 + k^2}. \tag{35}$$

If there were little gain or attenuation in propagation, q would be very nearly purely real, and would be substantially greater than k at high frequencies. The propagation constant γ would be nearly purely imaginary, and would have a magnitude nearly equal to q at high frequency.

We wish to determine γ and q when we include the negative conductance along the center, the spreading resistance, the skin resistance of the strip line and the positive conductance along the edges. These resistances will modify γ so that it will have a finite real part, indicating gain or loss for traveling waves. The method of attack is to find the admittance per unit length looking outward from the diode strip as a function of q and ω. This admittance must be the negative of the admittance per unit length of the diode strip, giving an equation for q as a function of frequency. This is based upon an argument that the current leaving the diode must enter the strip line, and, if a voltage is to exist, the admittances must be equal but of opposite signs. There is a hidden assumption here that the longitudinal current under the diode junction is negligible. This will be valid provided that the diode strip is *very* narrow compared with a wavelength on the structure and the strip-line gap is quite thin and of the same order of magnitude as the skin depth in the semiconductor under the junction. In this case, there will be a low-impedance current path from one part of the diode to another through the strip-line

gap and only a small fraction of the total z-directed current will be in the semiconductor material under the junction.

We can find the transverse admittance of the strip line by solving the boundary-value problem for the outer edge, temporarily neglecting skin-effect losses. At the outer edge we assume that the resistive film forms the total terminating admittance and that there is no y-directed current leaving the gap region. The wall resistivity at the boundary is taken as

$$\rho_s = \left(r\sqrt{\frac{\epsilon}{\mu}}\right)^{-1} \quad \text{ohms per square,} \tag{36}$$

where $1/r$ is the dimensionless ratio of the free-space impedance $\sqrt{\mu/\epsilon}$ to the actual resistivity ρ_s. At the boundary where $y = y_0$, therefore,

$$\left(\frac{H_z}{E_x}\right)_{y=y_0} = r\sqrt{\frac{\epsilon}{\mu}}. \tag{37}$$

Using (31) and (33), we can write the boundary equation as

$$\left(\frac{\gamma^2 + k^2}{(j\omega\mu q)}\right)\frac{Ae^{qy_0} + Be^{-qy_0}}{Ae^{qy_0} - Be^{-qy_0}} = r\sqrt{\frac{\epsilon}{\mu}}. \tag{38}$$

We may substitute for $(\gamma^2 + k^2)$ from (35) to obtain

$$\frac{\frac{A}{B}e^{2qy_0} + 1}{\frac{A}{B}e^{2qy_0} - 1} = -jr\frac{k}{q}, \tag{39}$$

from which we may solve for A/B, obtaining

$$\frac{A}{B} = e^{-2qy_0}\frac{jr\frac{k}{q} - 1}{jr\frac{k}{q} + 1}. \tag{40}$$

The admittance at $y = 0$ is obtainable by substituting the above and performing some algebra, giving

$$(Y_s)_{y=0} = \frac{1}{h}\left(\frac{H_z}{E_x}\right)_{y=0} = \frac{r}{h}\sqrt{\frac{\epsilon}{\mu}}\frac{1 - j\frac{q}{rk}\tanh qy_0}{1 + j\frac{rk}{q}\tanh qy_0} \quad \text{mhos per meter.} \tag{41}$$

We must also include the diode spreading resistance and the effect of skin resistance in the strip-line region. The spreading resistance under the diode can be taken as a long strip resistance with a conductivity

G_{s1} mhos per meter of length, and we may postulate another strip resistance with a conductance of G_{s2} mhos per meter to account for the skin resistance of the strip line. We include these terms to obtain the total admittance for the circuit looking outward from the diode junction:

$$Y_c = \frac{1}{\frac{1}{G_{s1}} + \frac{1}{2G_{s2}} + \frac{1}{2Y_s}} \quad \text{mhos per meter.} \tag{42}$$

The factors of 2 in the above account for the two sides of the strip line.

We can now write the admittance equation, which is solvable for q, the transverse propagation constant. This is

$$Y_d = -Y_c \tag{43}$$

or

$$\frac{1}{\frac{1}{G_{s1}} + \frac{1}{2G_{s2}} + \frac{1}{2Y_s}} = G_d - j\omega C_d\,, \tag{44}$$

where C_d and G_d are the capacitance and negative conductance per meter of length.

We need expressions for the terms G_{s1} and G_{s2}. The term G_{s1} is the inverse of the spreading resistance for one meter of diode length,

$$R_{s1} = \frac{1}{G_{s1}} = \frac{\rho}{\pi} \ln \frac{r_s}{r_d} \quad \text{ohm-meters,} \tag{45}$$

where ρ is the resistivity of the semiconductor material. Both G_{s2} and Y_s are transcendental functions of wavelength and frequency and involve the unknown q. We will have to make further simplifying assumptions to obtain an equation which we can use readily. For this, we assume that the strip line is sufficiently wide and the frequency is sufficiently high that $\tanh qy_0$ can be taken as equal to one. This is the same as taking A/B equal to zero in (31). This assumption is valid for the high-frequency region, where we can expect substantial gain and where the ac fields are weak at the edges of the strip line. This approximation gives

$$(Y_s)_{\text{high freq.}} = -j\frac{q}{h\omega\mu}\,. \tag{46}$$

To determine $1/G_{s2}$ we use the same high-frequency approximation. We also assume that skin effect perturbs the field equations but little, and compute the skin-effect power loss per unit length by an integral over the two faces of the strip line,

$$P = 2\int_0^{y_0(\to\infty)} R_{\mathrm{sk}}(H_y^2 + H_z^2)\,dy \qquad \text{watts per meter,} \tag{47}$$

where R_{sk} is the metallic skin resistance in ohms per square. We may note that H_z and H_y have nearly equal magnitudes at high frequency provided that $q^2 \gg k^2$. This allows us to integrate into a simple form:

$$P = \frac{2H_z^2(0)R_{\mathrm{sk}}}{\mathrm{Re}\,(q)} \qquad \text{watts per meter,} \tag{48}$$

where Re (q) is the real part of q. We postulate an equivalent strip resistance R_{s2} (or G_{s2}^{-1}) at the inside edge adjacent to the diode. This would give a power loss of

$$P = R_{s2}H_z^2(0) \qquad \text{watts per meter.} \tag{49}$$

Equating these powers gives an approximate value for R_{s2} :

$$\frac{1}{G_{s2}} = R_{s2} = \frac{2R_{\mathrm{sk}}}{\mathrm{Re}\,(q)} \qquad \text{ohm-meters.} \tag{50}$$

Let us now write (44) in a complete form valid at high frequencies only:

$$\frac{1}{\dfrac{\rho}{\pi}\ln\dfrac{r_s}{r_d} + \dfrac{R_{\mathrm{sk}}}{\mathrm{Re}\,(q)} + j\dfrac{h\omega\mu}{2q}} = G_d - j\omega C_d\,. \tag{51}$$

We may solve this for q by using still another approximation — that q is substantially a real quantity — and substitute q for Re (q). This is a valid approximation because the term involving Re (q) is small compared with the one involving q:

$$q = \frac{\omega^2 C_d \mu h}{2}\left[\frac{\left(1 - j\dfrac{2R_{\mathrm{sk}}}{\omega\mu h}\right)\left(1 + \dfrac{jG_d}{\omega C_d}\right)}{1 - \left(\dfrac{\rho}{\pi}\ln\dfrac{r_s}{r_d}\right)G_d + j\left(\dfrac{\rho}{\pi}\ln\dfrac{r_s}{r_d}\right)\omega C}\right]. \tag{52}$$

At high frequencies, the imaginary quantities in the numerator of (52) are small compared with one, and with low spreading resistance the second and third terms in the denominator are also small compared with one. If we bring up the denominator with a binomial expansion, multiply the result and drop terms involving the products of two or more small quantities, we obtain an approximate first-order result:

$$q \approx \frac{\omega^2 C_d \mu h}{2}\left(1 + \frac{jG_d}{\omega C} - j\frac{2R_{\mathrm{sk}}}{\omega\mu h} - j\frac{\rho}{\pi}\left(\ln\frac{r_s}{r_d}\right)\omega C_d + \frac{\rho}{\pi}G_d\ln\frac{r_s}{r_d}\right). \tag{53}$$

From q we can obtain γ by (35). This is

$$\gamma = \pm j\sqrt{q^2 + k^2}. \tag{54}$$

In the high-frequency region where (46) is valid, γ^2 will be much larger than k^2, so that

$$\gamma \approx \pm jq. \tag{55}$$

We will obtain gain if the real and imaginary parts of γ have opposite signs, and this requires a positive imaginary part in the expression for q given in (53). The maximum frequency for this is

$$f_{\max} = \frac{G_d}{2\pi C_d}\sqrt{\frac{1 - \dfrac{2R_{sk}C_d}{\mu h G_d}}{G_d \dfrac{\rho}{\pi} \ln \dfrac{r_s}{r_d}}}. \tag{56}$$

If we ignore skin-effect losses, this is the same limit we obtained in (20). This can easily be shown by substituting $G_d = \pi r_d G_{d1}$, $C_1 = \pi r_d C_{d1}$ and $R_{sk} = 0$.

There are several drawbacks to this structure as a practical amplifier. The major one is that the circuit will show gain in both directions and there will be considerable difficulty in stabilization. Slight mismatches will establish standing waves, and oscillations can occur if the total gain is substantial. Gain will be found over a wide frequency range, and it will be difficult to obtain a good match to input and output transmission lines throughout this range. The impedance of the structure is approximately proportional to frequency over a wide range, and this increases the matching difficulties.

A method of avoiding oscillations is to combine this circuit with a nonreciprocal attenuator to give a net loss in one direction and permit gain in the other. One method of accomplishing this is illustrated in Fig. 17. (This proposal was made in collaboration with W. W. Anderson of Bell Telephone Laboratories.) It involves the placement of thin ferrite slabs in both sides of the strip-line gap. The ferrite material is magnetized approximately in the $\pm x$ direction in a nonuniform manner by the shaping and placement of the magnetic pole pieces. This field is strong near the center and weaker near the edges of the strip line. This provides for a region of high-frequency magnetic resonance near the diode strip, continuously graded to a lower frequency resonance at greater distances from the center. It will be noted in (31) and (33) that the ratio of H_y to H_z at high frequency is

$$\frac{H_y}{H_z} = \pm j\frac{\gamma}{\sqrt{\gamma^2 + k^2}} \tag{57}$$

provided that A/B can be taken as approximately zero. This indicates that the magnetic field will be elliptically polarized. If the wave velocity is so slow so that $\gamma^2 \gg k^2$, the polarization will be quite nearly circular. The ellipticity will be pronounced only at low frequencies, where the wave velocity is more nearly the velocity of light, and near the outer boundary, where Ae^{qy} may be roughly equal to Be^{-qy}. As in other devices of this type, the sense of rotation of the magnetic vector is reversed if the direction of propagation is reversed. This is the condition for nonreciprocal resonance absorbtion in ferrite materials that are magnetized perpendicular to the plane of polarization.

A unique property of this circuit is that the circularity condition is met over a wide frequency range and also over a large fraction of the available space. This allows us to obtain broadband isolation by using a nonuniform magnetization for the ferrite isolator, so that resonance for the highest frequencies occurs near the center and, at lower frequencies, at greater distances from the axis. The steady H_x field is to be nonuniform and vary roughly as y^{-1}. This requires that H_y be nonzero if Laplaces's equation is to be satisfied. However, if the slab is thin, the H_y component will be small and give little trouble.

A complete description and theory for this type of isolator is beyond the scope of this paper. The author is not aware of any theoretical solu-

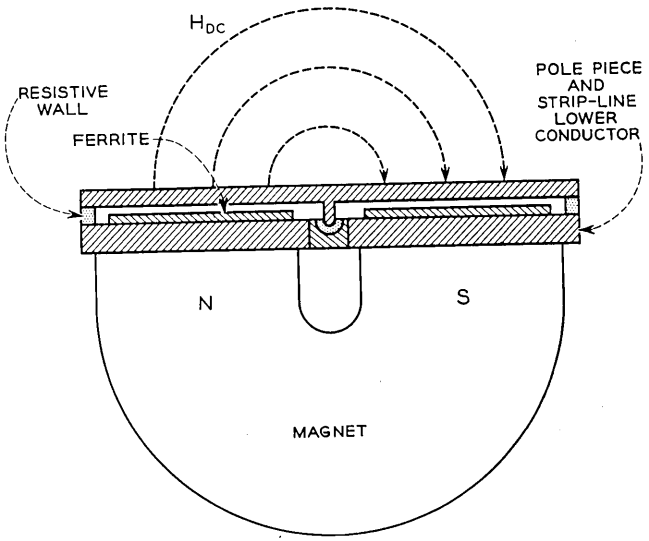


Fig. 17 — A proposed method of obtaining unidirectional gain for the circuit of Fig. 15. The ferrite acts as a broadband resonance isolator distributed along the circuit, introducing substantial attenuation for only one direction of wave propagation.

tion to the wave propagation problem in a gyromagnetic medium that is nonuniformly magnetized. However, initial experiments by W. W. Anderson and the author are most encouraging. An isolation tester was built using a closely-spaced array of passive capacitors of 7 $\mu\mu$f each, spaced 0.065 inch apart along the axis of a thin waveguide 2 inches wide and 0.025 inch high. This gave an upper cutoff frequency as a filter at about 8 kmc and a lower cutoff due to the metallic outer boundary at about 1 kmc. Through most of the intervening band, the wave propagation characteristics were similar to the continuously loaded strip line we have been discussing. The insertion of a single type of ferrite material into the 0.025-inch space on both side of the gap gave a maximum forward additional attenuation of 0.5 db per inch and a minimum additional reverse attenuation of 15 db per inch over the band from 1.5 kmc to 6.0 kmc. These measurements were made with a single setting of the magnetic field.

It is expected that the first practical traveling-wave amplifier of this type may involve the use of a closely spaced array of spot diodes rather than a continuous strip diode. If the gain per unit length is to be sufficiently low that the ferrite material can suppress reverse gain, it may be necessary to use relatively few active diodes, interspersed with passive capacitors to keep the wave velocity low.

VI. STRIP-DIODE OSCILLATOR CIRCUITS

The simplest method of obtaining oscillations with a narrow-strip diode might be to use a short section of a circuit like that of Fig. 15. The length ΔZ should be equal to one-half of a wavelength on the circuit at the desired frequency. In this case, the circuit should be shorter than one free-space wavelength at the *maximum frequency of amplification* as given by (56) if we desire absolute stability in the next-order mode of resonance. Probably, however, this limit can be exceeded to some extent in a practical oscillator, as nonlinearities are sometimes effective in allowing only one mode of oscillation at full power. Fig. 18 shows such a circuit, which can be mounted in a waveguide in a manner similar to the circuit of Fig. 13.

Fig. 19 shows a possible arrangement utilizing two strip diodes, intended to operate in a push-pull mode. This is still closer in concept to the circuit of Fig. 13 and could be mounted in the same way. It should operate at substantially higher frequencies than that of Fig. 18, and allow a greater total diode area. In this case, the current paths in the desired mode are shorter, reducing the effects of skin-effect resistance in the circuit.

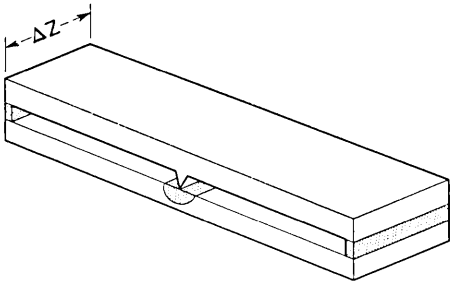


Fig. 18 — A possible high-frequency oscillator circuit using a narrow-line Esaki diode. This is simply a short section of the circuit of Fig. 15. The length ΔZ is one-half wavelength at the desired frequency. This could be mounted in the same way as the circuit of Fig. 13.

In the desired mode of operation, each of the two diode strips should have uniform phase along its length, but the two should be 180° out of phase with respect to each other. The slot between the diodes acts as an inductive chamber to resonate the capacitance of the diodes in series. We must be concerned also with three spurious modes if we desire to use the maximum possible diode strip length. The four lowest-order modes are illustrated in Fig. 20. Here the surface current flow is shown on one internal face of the strip line for each mode. In the desired push-pull mode, substantially all of the current flows directly across the inductive slot between the strips, but some small amount flows outward to charge the capacitance between the plates on the "land" regions on either side of the strips. Transverse slots are placed at each end of the strips to provide an inductive impedance for currents attempting end-runs around

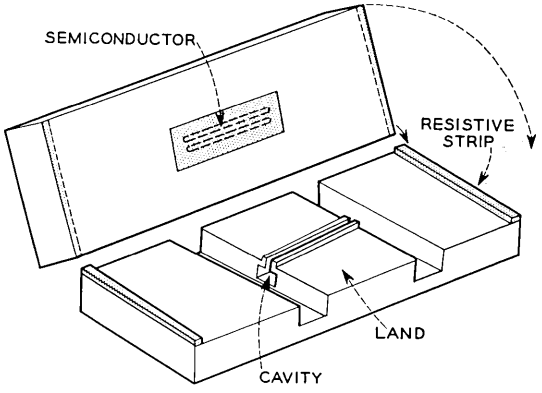


Fig. 19 — A possible push-pull circuit using two narrow-line Esaki diodes.

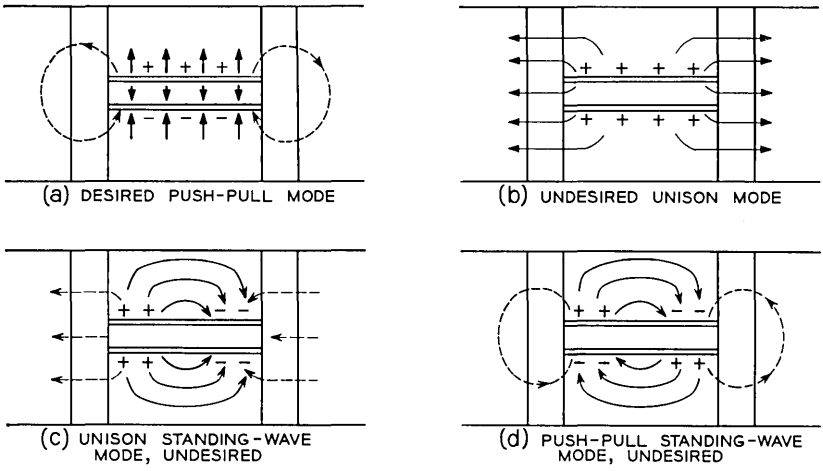


Fig. 20 — Four possible modes of oscillation for the circuit of Fig. 19. The surface current flow is illustrated for the semiconductor side of the strip line. The resistance strips can suppress the unison mode of (b). The standing-wave modes of (c) and (d) can be suppressed by sufficiently short strips and using close gap spacing in the "land" region of Fig. 19.

the ends of the strips from one side to the other. In the undesired unison mode the two diodes are in equal phase and this mode would launch waves along the strip lines, which must have very narrow gaps and low impedances if this mode is to be suppressed. The two other undesired modes involve standing waves along the diode junctions. In one, the opposing diode strips are in phase, and in the other they are 180° out of phase.

The problems of analysis will be to determine the frequency for the desired mode, to determine whether or not the desired oscillation will occur, and to determine whether or not oscillations in the three listed spurious modes will be troublesome. Of course, still higher-order standing-wave modes are possible, but these will be at still higher frequencies and therefore less troublesome.

For the desired mode, we can determine the inductance of the central cavity for a given length. At the desired frequency this inductive reactance should equal that of the capacitance of the two diodes in series. If the "lands" have appreciable capacitances, these must be added to the diode capacitances. The methods of accounting for spreading resistance and circuit resistance are similar to that used for the traveling-wave amplifier described in Section V.

For the undesired standing-wave modes we can determine the resonant

frequencies by first considering that the circuit extends indefinitely in the direction of the diode strips. We can find two propagation characteristics of such a structure, one for the diodes operating in unison and one for push-pull phase opposition. This can be done by an extension of the methods of the last section, taking proper account of the central cavity. The resonant frequencies for a limited diode length will be those frequencies for which the diode strips are a half-wavelength in extent. These frequencies must be above the maximum frequency for gain as a traveling-wave device if there is to be no danger of oscillation in spurious modes. This condition will establish a maximum length for the diode strips. If the land regions have a very narrow gap, this will allow longer strips but will also increase the circuit losses. A compromise can be reached, however, which will allow a substantial diode length with only slight degradation of the desired operation.

This type of circuit has been analyzed rather completely in unpublished work by the author, but the theory is too lengthy to include in this paper. A particular fictitious example to operate at about 9000 mc would require an R_nC product of 10^{-10}, diode strips only 0.00016 inch wide and strip-line and land gaps of 0.00016 inch. The resonant cavity required would be approximately 0.003 $\times$ 0.003 inch in cross section! However, diode strips on the order of 3 mm long would be allowable without danger of standing-wave oscillations of the types shown in Fig. 20. For this example, the total diode direct current would be about 300 ma, and the power output might be a few milliwatts, compared to the few microwatts that could be obtained from a single-spot diode small enough to oscillate at this frequency. It is obvious that a practical utilization of this principle will involve some difficult fabrication problems.

A third method of using a narrow strip diode as an oscillator is to use an axially symmetric geometry with a ring-shaped diode. The structure is illustrated in Fig. 21. The desired mode of operation involves no angular variations of voltage around the ring, and the cavity is an annular space adjacent to the ring. Outside of the cavity, there is a large bypass capacitance.

In this structure, the ring diameter would be limited by the possibility of oscillations in modes involving angular variations of voltage around the ring. To assist in obtaining a larger ring diameter we can use a very narrow gap spacing in the region inside the ring. In a manner analogous to the action of the lands in the two-strip structure, this space acts as a low inductance transverse current path for standing-wave patterns around the ring, raising their resonant frequencies and inducing extra loss. In the desired mode, the central cavity region acts as an addi-

tional capacitance to be added to that of the diode. Current paths for the lowest-order mode of this type are shown in Fig. 21.

Methods of analysis for this structure in the desired mode are quite straightforward. The methods outlined in previous sections of this paper should be adequate. For the undesired θ-varying modes, we can use a method analogous to that used for the traveling-wave amplifier. A complete analysis will not be given here, but a method of attack will be described.

As before, we can determine the admittance per unit length along the diode ring looking inward and outward. Looking outward, we simply have the inductive reactance of the cavity plus the circuit losses. Looking inward, we have a simple capacitance for the desired unison mode. For the lowest-order standing-wave mode we can write the appropriate field equations for the space, presuming that E_z varies as

$$E_z = AJ_1(k_c r) \cos \theta, \tag{58}$$

from which we can determine H_r and H_θ.

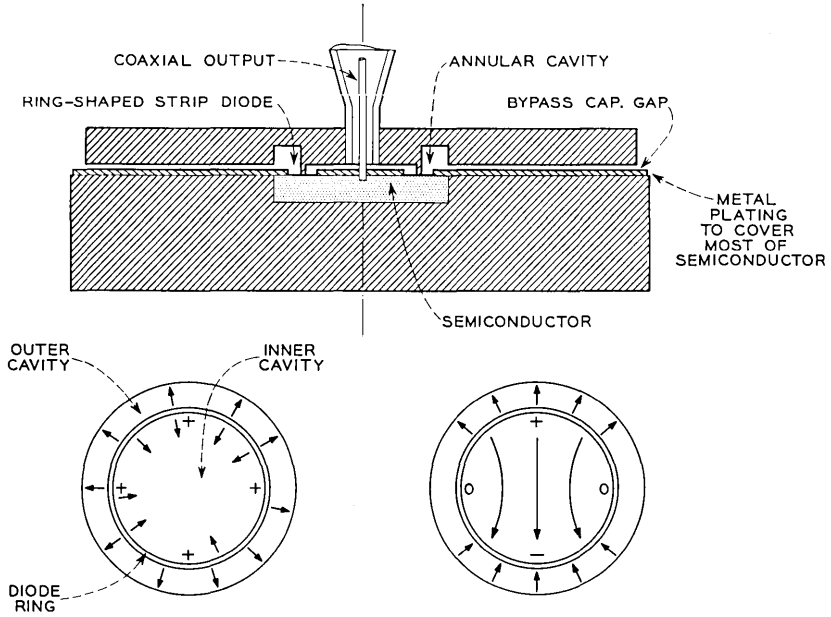


Fig. 21 — A possible oscillator circuit using a narrow-line Esaki diode in ring form. The desired mode involves uniform phase around the ring, as shown at the lower left. The undesired modes will involve θ-variations, as shown at the lower right. The latter can be suppressed by keeping the ring diameter sufficiently small and using a narrow gap in the inner cavity.

This will give the reactance for a unit length along the circumference,

$$jX = \frac{hE_2}{H_\theta} \qquad \text{ohm-meters,} \tag{59}$$

which will be inductive if the inner cavity is small compared to a wavelength.

The ring structure and the two-strip structure are quite similar circuits in many ways. The ring structure has a disadvantage in that it will be more difficult to obtain low circuit losses at the highest frequencies in the desired mode. The extra circuit losses will be found in the resonant cavity because it must have a larger ratio of perimeter to cross-section area, in the semiconductor material because the gap is not shared by two diodes in series, and in the bypass condenser that is avoided in the two-strip structure. It may also prove to be more difficult to arrange suitable coupling to a waveguide or coaxial line output for the ring structure. The coaxial output shown does not appear to be an entirely satisfactory answer.

VII. NOISE FIGURE FOR A SMOOTH DISTRIBUTED AMPLIFIER

In this section we will derive an expression for the noise figure of a rather generalized type of distributed amplifier in which isolation is not used. Chang[9] and Anderson and Hines[11] have derived expressions for the single-stage amplifier. We assume that we have a transmission line with a series reactance jX_1 and a series passive resistance of R_1 ohms per meter. The gain is provided by a distributed negative conductance of $-G_d$ mhos per meter, and we have an additional passive shunt conductance G_1 and a susceptance of jB_1 mhos per meter. We may use the well-known equation for such a line, which states that waves propagate as

$$\begin{aligned} V(z,t) &= V(0,0)e^{j\omega t-\sqrt{ZY}\,z} \\ &= V(0,0)e^{j\omega t-j\beta z+\alpha z}, \end{aligned} \tag{60}$$

$$Z = R_1 + jX_1\,, \tag{61}$$

$$Y = G_d + G_1 + jB_1\,. \tag{62}$$

The characteristic impedance is given by

$$Z_0 = R_0 + jX_0 = \sqrt{\frac{Z}{Y}} = \frac{V(z,t)}{i(z,t)}\,. \tag{63}$$

The noise figure of an amplifier system depends upon the termination

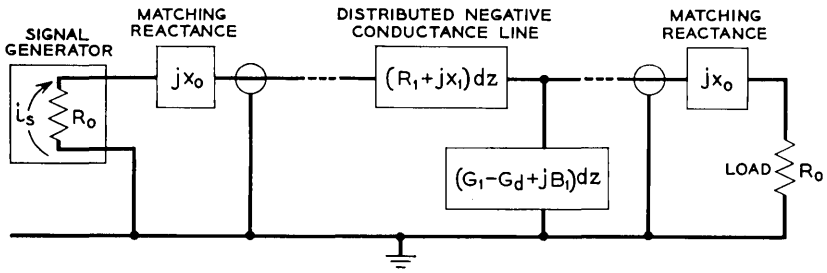


Fig. 22 — A generalized type of negative-conductance traveling-wave transmission line. Expressions have been derived for the noise figure of such a line.

presented at its source and in some cases at the load also. Ordinarily, matched impedances are considered desirable. In this case, the impedance of the amplifier line is complex, which presents a special problem, since we cannot match the amplifier impedance to the input and output lines for both directions of propagation. To prevent internal standing waves, a line with characteristic impedance $R_0 + jX_0$ should see an impedance $R_0 + jX_0$ at either end. In a conventional network, we would "match" by terminating with $R_0 - jX_0$ to obtain maximum power transfer; here, however, we wish to terminate with the characteristic impedance rather than its complex conjugate. Let us terminate our lines in this way, as shown in Fig. 22. This will cause some power reflection at the input but will avoid internal standing waves and undesired regeneration effects.

A current i which enters the amplifier grows with distance as

$$i(z) = i_1 e^{\sqrt{zY}\, z} = i_1 e^{-j\beta z + \alpha z}. \tag{64}$$

At the output

$$i_2 = i_1 e^{\alpha l}, \tag{65}$$

where l is the total length. We define the "internal gain" g_i as

$$g_i = e^{2\alpha l}. \tag{66}$$

For the circuit of Fig. 22, the rms current source i_s for a sine-wave input of available power p_i is

$$i_s = 2\sqrt{\frac{p_i}{R_0}}, \tag{67}$$

which divides between R_0 and the input in proportion to the relative admittances. The current i_1 is easily determined to be

$$i_1 = \sqrt{\frac{p_i}{R_0}} \frac{R_0}{\sqrt{R_0^2 + X_0^2}}. \tag{68}$$

The output power is $i_2^2 R_0$, which determines the net gain g_p,

$$g_p = g_i \frac{R_0^2}{R_0^2 + X_0^2}. \tag{69}$$

We will use the following expression as our definition of noise figure F:

$$F = 1 + \frac{\text{output noise power from internal sources}}{(\text{net gain})\ (kT_0B)}, \tag{70}$$

where k is Boltzmann's constant, T_0 is the noise reference temperature (290°K) and B is the bandwidth of interest. Internal sources include the shot noise of the diode direct current plus thermal noise in the internal passive resistance. We assume a direct current in the diode of I_0 amperes per meter. For the diode noise we might assume shot noise or something proportional to shot noise. In an infinitesimal length dz, we can assume a differential amount of mean-square noise current,

$$d(\bar{i}^2) = \gamma^2 2eI_0B\, dz \qquad \text{amperes}^2, \tag{71}$$

where γ^2 is our unknown factor of proportionality and e is the electronic charge. This current from an infinitesimal section of line must divide equally into a forward and a backward wave, with the backward wave presumed to be lost. The forward waves from each length dz add at the output with an appropriate gain factor depending upon the distance from the point z and the output. We can find the total mean-square current at $z = l$ by a simple integration:

$$\begin{aligned}(\bar{i}_2^2)_{\text{shot noise}} &= \frac{\gamma^2 eI_0B}{2} \int_0^l e^{2\alpha z}\, dz \\ &= \left(\frac{\gamma^2 eI_0B}{2}\right)\left(\frac{g_i - 1}{2\alpha}\right) \qquad \text{amperes}^2.\end{aligned} \tag{72}$$

We can find the total mean-square current at $z = l$ for the thermal noise in the shunt conductance G_1 in the same manner. For an infinitesimal conductance, $G_1 dz$, the noise can be considered to arise in a current generator,

$$d(\bar{i}^2)_{G_1} = 4kTBG_1\, dz, \tag{73}$$

and we integrate as before to obtain

$$(\bar{\imath}_2^2)_{G_1} = kTBG_1 \frac{g_i - 1}{2\alpha} \quad \text{amperes}^2. \tag{74}$$

For the series resistance R_1 we assume the alternative Thevenin form of a voltage source of noise for a length dz:

$$d(\bar{v}^2)_{R_1} = 4kTBR_1\,dz. \tag{75}$$

There is an impedance $2(R_0 + jX_0)$ in series with this generator, giving

$$d(\bar{\imath}^2)_{R_1} = \frac{kTBR_1}{R_0^2 + X_0^2}\,dz, \tag{76}$$

which we integrate as before to obtain

$$(\bar{\imath}_2^2)_{R_1} = \left(\frac{kTBR_1}{R_0^2 + X_0^2}\right)\left(\frac{g_i - 1}{2\alpha}\right). \tag{77}$$

To obtain the total output noise power from internal sources we add the three mean-square currents and multiply by R_0. From (70) we obtain

$$F = 1 + \left(\frac{g_i - 1}{g_i}\right)\left(\frac{R_0^2 + X_0^2}{R_0^2}\right)\left(\frac{T_d}{T_0}\right)\left(\frac{R_0 G_d}{2\alpha}\right) \cdot \left(\frac{\gamma^2 e I_0}{2kT_d G_d} + \frac{R_1}{G_d(R_0^2 + X_0^2)} + \frac{G_1}{G_d}\right). \tag{78}$$

It is evident that we want R_1 and G_1 to be small. These are parasitic elements which are generally undesirable. We would also like to reduce the magnitude of the term $eI_0/2kT_0G_d$. This term depends upon the properties of the p-n junction itself and is independent of junction area or circuitry. To obtain the lowest noise figure we should seek a diode which has a high negative conductivity per unit of direct current.

Equation (78) can be simplified for the limiting case of no internal passive resistance, that is for $R_1 = 0$ and $G_1 = 0$. This allows us to drop two of the three terms in the last factor. Also, the terms $(R_0^2 + X_0^2)/R_0^2$ and $R_0G_d/2\alpha$ are reciprocals in this special case and cancel.

This leaves

$$(F)_{\substack{R_1=0 \\ G_1=0}} = 1 + \left(\frac{g_i - 1}{g_i}\right)\left(\frac{\gamma^2 e I_0}{2kT_0G_d}\right). \tag{79}$$

This is essentially the same expression obtained by Hines and Anderson[11] for the case of a single diode amplifier without parasitic resistance.

VIII. GENERAL DISCUSSION AND CONCLUSIONS

In this paper at least some of the methods have been described by which we may hope to obtain appreciable amounts of power in the micro-

wave range with circuits using Esaki diodes. The circuits proposed would utilize diodes in pairs, in narrow strip form, and in traveling-wave distributed circuits with ferrite nonreciprocal attenuation.

The situation appears very hopeful to the author. As our semiconductor technology improves, we should be able to develop useful solid state amplifiers and oscillators for the high microwave and probably for the millimeter-wave range as well.

The author suspects that solid state device research for the microwave and millimeter-wave range will probably continue to advance through the discovery of better methods of obtaining negative-resistance effects. Esaki diodes are usable devices at high frequencies, and we must exploit their possibilities. What we learn in the process about the useful application of negative resistance will probably be helpful in designing better devices to come.

REFERENCES

1. Esaki, L., Phys. Rev., **109,** 1958, p. 603.
2. Sommers, H., Proc. I.R.E., **47,** 1959, p. 1201.
3. Rutz, R. F., I.B.M. J. Res. Dev., **3,** 1959, p. 372.
4. Hall, R. N., I.R.E. Prof. Group on Electron Devices, Washington, D. C., October 1959.
5. Batdorf, R. L., Dacey, G. C., and Wallace, R. L., to be published.
6. Holonyak, N., Jr., Lesk, I. A., Hall, R. N., Tiemann, J. J., and Ehrenreich, H., Phys. Rev. Letters, **3,** 1959, p. 167.
7. Chynoweth, A. G., Feldmann, W. L., Lee, C. A., Logan, R. A., Pearson, G. L. and Aignain, P., to be published.
8. Thomas, D. E., U. S. Patent No. 2,896,168, July 21, 1959.
9. Chang, K. K. N., Proc. I.R.E., **47,** 1959, p. 1268.
10. Seidel, H. and Herrmann, G. F., I.R.E. Wescon Conv. Rec., 1959, Part 2, p. 83.
11. Hines, M. E. and Anderson, W. W., Proc. I.R.E., **48,** 1960, p. 789.

Theory of Current-Carrier Transport and Photoconductivity in Semiconductors with Trapping

By W. VAN ROOSBROECK

(Manuscript received September 30, 1959)

Fundamental differential equations are derived under the unrestricted approximation of electrical neutrality that admits trapping. Extension is made for applied magnetic field. The transport equations derived hold without explicit reference to detailed trapping and recombination statistics. Modified ambipolar diffusivity, drift velocity and lifetime function apply in the steady state. The same diffusion length is shown to hold for both carriers, and a general "diffusion-length lifetime" is defined. Mass-action statistics are considered for cases of (one or) two energy levels. Certain "effective" — rather than physically proper — electron and hole capture and release frequencies or times that apply to concentration increments are defined. Criteria are given for minority-carrier trapping, recombination and majority-carrier trapping, and for "shallow" and "deep" traps. Applications of the formulation include: the diffusion-length lifetime for the Shockley-Read electron and hole lifetimes; linear and nonlinear steady-state and transient photoconductivity; negative photoconductivity; the photoconductive decay observed by Hornbeck and Haynes in p-type silicon; the photomagnetoelectric effect; and drift of an injected pulse. Photomagnetoelectric current is found to be decreased by minority-carrier trapping, through an increase in diffusion length. A simple general criterion is given for the local direction of drift of a concentration disturbance. With trapping, there may be "reverse drift," whose direction is normally that for the opposite conductivity type, and also local regions of carrier depletion that may extend in practice over appreciable distances.

TABLE OF CONTENTS

I. Introduction and Outline 516
1.1 Introduction 516
1.2 Outline of Procedures and Results 518
1.3 List of Symbols 528

II. General Formulation 533
2.1 The Transport Equations 533
2.1.1 Extension for Applied Magnetic Field 537
2.1.2 Formulation for the Steady State in Terms of Trapping Ratios. 539
2.2 Mass-Action Theory 542
2.2.1 Single-Level Centers of Two Types 542
2.2.2 Centers with Two Energy Levels 552
2.2.3 Volume Generation with Excitations Involving Trapping Levels 554
III. Detailed Theory and Applications 554
3.1 Diffusion Length and Steady-State Lifetime Functions 554
3.1.1 Linear Theory 554
3.1.2 Nonlinear Theory 559
3.2 Photoconductivity 564
3.2.1 Linear Theory 564
3.2.2 Nonlinear Theory 569
3.2.3 Negative Photoconductivity 574
3.2.4 Further Theory with an Application to Experiment 575
3.3 The Photomagnetoelectric Effect 582
3.4 Transport of Injected Carriers 588
3.4.1 The Linear Differential Equations 588
3.4.2 Steady-State Transport; Reverse Drift 589
3.4.3 Drift of an Injected Pulse 591
IV. Acknowledgments 605
Appendix A. Derivation of Two-Sided Laplace Transforms 606
Appendix B. Solutions for Drift of an Injected Pulse 608
Appendix C. Integrals Over the Drift Range 609
References 611

I. INTRODUCTION AND OUTLINE

1.1 *Introduction*

Implications of the negligible space charge generally associated with carrier injection in homogeneous semiconductors have been worked out in some detail in phenomenological transport theory. However, the particular condition of electrical neutrality employed — that of constant excess of one mobile-carrier concentration over the other — is a restricted one that applies as an approximation in some cases. Upon injection, changes generally occur in concentrations of fixed charges associated with various impurities or crystal imperfections, including those on which equilibrium conductivity and those on which equilibrium lifetime, as a rule, largely depend. In a general sense, these concentration changes constitute *trapping*.

The literature on trapping is extensive,† and it is recognized that the subject entails difficulties that seem at odds with fairly simple essentials. The four processes of capture and release of electrons and holes for each type of center proceed at rates that depend on the concentration, energy level and capture cross sections of the centers, and also on concentrations

† Some general discussions of trapping are given in chapters of Breckenridge, Russell and Hahn,[1] Hoffman,[2] Fan[3] and Shulman.[4] The last includes many references as does the review of Bemski.[5]

of trapped and mobile excess carriers and of carriers at thermal equilibrium. A given general model thus presents a variety of physical possibilities, especially if multilevel centers or centers of more than one type are involved. Centers at given temperature may, for example, give mostly trapping or mostly recombination, depending on conductivity and conductivity type.† Moreover, as will be shown, small-signal nonlinearity with markedly nonconstant lifetimes usually occurs with minority-carrier trapping unless the concentrations of centers or of added electrons and holes are small compared with a concentration of the order of the minority-carrier concentration at thermal equilibrium. Nonconstant lifetimes may occur in the transition to the large-signal range as well. With trapping, familiar approximations relating only to the effect on conductivity are not particularly useful, and the more easily obtained approximate solutions very frequently do not apply. As these considerations indicate, a general treatment is necessary. This paper‡ gives ambipolar theory based on the unrestricted neutrality condition, with applications to problems in transport and photoconductivity, including a specific application to experiment. It extends results previously reported[9] and places some emphasis on: rigorous phenomenological formulation; classification with respect to types of physical behavior; conditions for the validity of approximations; techniques for analysis of given models; determination of capture cross sections and energy levels; and illustrative cases, which, through identification of the physical processes, provide qualitative insight.

In Section II, the first of two main sections, fundamental differential equations are derived that take into account diffusion, drift, recombination and trapping. This section also includes: a specialization to the steady state, which exhibits how trapping (of arbitrary statistics) modifies recombination and the transport processes; definitions of certain "effective" frequencies and times that properly characterize trapping and recombination as they apply to concentration increments above thermal equilibrium; certain fundamental relations from detailed balance; and criteria for classifying centers with respect to their trapping and recombination properties.

In Section III, the second main section, the general ambipolar formulation is applied to investigate trapping in various connections. From the theory for the steady state, diffusion lengths and lifetime functions are evaluated, and the photomagnetoelectric effect is analyzed. Transient

† With change in temperature, capture cross sections (as well as conductivity) may change appreciably. See Shulman,[4] Bonch-Bruevich[6] and Sandiford.[7]

‡ It is expected that a supplementary abridgement of the present paper will be published.[8]

photoconductivity also is analyzed critically and in some detail, a procedure facilitated by results from the present formulation of comparative formal simplicity. This analysis involves a formalism that recurs in theory of time-dependent transport. A detailed treatment of the drift with trapping of an injected pulse is given. These applications of the formulation constitute an illustrative selection.

In Section 1.2 is assembled descriptive material intended to be read for further preliminary orientation as to the contents of the paper, and also to be read piecemeal with corresponding portions of the main sections.

1.2 *Outline of Procedures and Results*

The formulation is accomplished in two stages. By treating concentrations of added electrons and holes formally as unrelated variables, differential equations for the transport are derived in Section 2.1 along the lines of previous treatments.[10,11]† Extension for applied magnetic field is included.[11] These equations involve no specific reference to the detailed trapping and recombination statistics. Specialized to the steady state, the ambipolar continuity equation is formally the no-trapping equation, but with the sum of fixed and mobile positive (or negative) charges as dependent variable, and with suitably modified ambipolar diffusivity, drift velocity, and lifetime function, which depend in general on two (concentration-dependent) phenomenological differential "trapping ratios." The same diffusion length is shown to apply for both electrons and holes, and a general "diffusion-length lifetime," τ_0, based on the unmodified ambipolar diffusivity, is defined. The formulation is completed in Section 2.2 with equations for the time rates of change of concentrations of carriers trapped in centers of each type.

These rates are written in accordance with mass action, which provides a simple‡ and general§ basis for trapping and recombination.‖ Two

† In Ref. 11 small Hall angles are assumed, in part because appreciable magnetoresistance is otherwise involved. As indicated in this reference, arbitrary Hall angles (and injection levels) could suitably be taken into account by theory involving the phenomenological magnetoresistance without added carriers.

‡ See Hoffmann.[2] The mass-action approach, now widely used in semiconductor theory, is essentially that used in early theory of metal-semiconductor junctions: see Schottky and Spenke.[12]

§ Boltzmann statistics, assumed for the transport equations, imply mass-action relationships at equilibrium: see Spenke.[13] But, with definitions of equilibrium parameters suitably extended, mass-action equations apply also for degenerate semiconductors: see Rose.[14]

‖ A treatment based on Fermi statistics that allows for degeneracy and includes dependence of occupation probabilities on applied magnetic field has been given by Landsberg.[15]

energy levels (as well as a single one) are considered; equations are written in Section 2.2.1 for two types of trapping centers and, through simple formal modification, in Section 2.2.2 for two levels from centers of a single type.† Partly by way of notational convention, the levels are taken as acceptor and donor levels, which give negative and positive fixed charges. This case is the simplest for which both steady-state trapping ratios occur, these ratios being the respective changes in concentration of all fixed negative charges and all fixed positive charges divided by the change in concentration of all negative or positive charges. With suitable interpretation of the notation, the equations apply to one- or two-level cases in general; results written for centers of the acceptor type, for example, are not restricted to this type. Moreover, it will appear that, in the analysis of transient (or steady-state) photoconductivity for a given multilevel model, the trapping at a given time need usually be considered in detail in no more than two successive levels. Levels appreciably lower and higher than these may contribute to recombination, but will not contribute to trapping, in the sense that the lower levels may be assumed to remain completely full (or else saturated) and the higher levels completely empty.‡

To facilitate analysis and interpretations, in Section 2.2.1.2 "effective" capture and release frequencies and times that apply to concentration increments are defined *a priori* from the mass-action equations. The four effective frequencies or times for each energy level differ from the physically proper ones, which depend on the trapped concentrations and thus on the detailed solution of the particular problem. They satisfy a fundamental restriction, used extensively in the theory, which is derived from thermal-equilibrium relationships involving detailed balance. With this restriction, quantitative criteria are established in Section 2.2.1.3 for ranges of minority-carrier trapping, recombination and majority-carrier trapping. These ranges may be specified in terms of the location of the equality level§ relative to the Fermi level $\mathcal{E}_F$ and the "reflected Fermi level" $\mathcal{E}_F'$, the reflection of $\mathcal{E}_F$ about $\mathcal{E}$, its location for intrinsic material. If spins are taken into account, quantities of the mass-action theory serve to locate the trapping level relative to $\mathcal{E}$. It is shown

† Theory for multilevel centers is given in Landsberg,[15,16] Champness,[17] Okada,[18] Shockley and Last,[19] Mercouroff,[20] Khartsiev,[21] Sah and Shockley,[22] Bernard,[23] Kalashnikov and Tissen,[24] and Kalashnikov.[25]

‡ The influence of trapping at a given level on recombination at another has been calculated for the near-equilibrium steady state by Kalashnikov.[26] See also Mashovets.[27]

§ This is the Fermi level for which the (equilibrium) rates of electron and hole capture and release are all equal.[22,28] The equality level is similar in purport to the demarcation level of Rose,[29,30] which is the trapping level for which the rates are equal.

that a proper criterion for "shallow" or "deep" minority-carrier trapping levels is that $\mathcal{E}_F'$ separates these levels. Thus, levels in extrinsic material considerably shallower than the midgap may still be "deep" levels.

Detailed theory is given in Section III through various applications of the general formulation, and consequences of the mass-action statistics are examined. In Section 3.1.1 diffusion length and diffusion-length lifetime, as well as the trapping ratios, are evaluated from equations written for the limiting linear small-signal steady state. A "capture concentration" is introduced, use of which is found to simplify formally much of the detailed theory, including that for time-dependent cases. This concentration is the concentration of (single-level) centers multiplied by the respective equilibrium fractions of centers occupied and unoccupied. Values of it that are small or large result, respectively, in negligible capture frequencies or in large capture frequencies with negligible release frequencies. For the case of a single energy level, the general (equilibrium) Shockley-Read electron and hole lifetimes[31] are obtained in forms involving the capture concentration. These lifetimes are shown to correspond to a diffusion-length lifetime τ_0 whose general expression is formally the same as that for the common (equilibrium) lifetime[31,32] in the limit of small concentration of centers. This common lifetime otherwise applies as such only under a condition restricting the capture concentration, which is frequently severe: In the minority-carrier trapping range, it is that this concentration be small compared with the equilibrium minority-carrier concentration. From conditions for the neglect of quadratic terms in the mass-action equations, the linear approximation is shown to imply a restriction of injection level that may be much more severe than the familiar small-signal condition[10] based on the conductivity change.

The general single-level trapping ratios and lifetime functions for the nonlinear steady state are obtained in Section 3.1.2. These and the mobile-carrier concentrations, as well as the volume recombination rate, can be expressed in terms of trapped-carrier concentration as single concentration variable. The lifetime functions reduce to the Shockley-Read lifetimes in the linear small-signal limit and to a single limiting large-signal value. The familiar common lifetime function[31] for small concentration of centers usually does not apply in the small-signal range unless it is substantially constant in this range. The differing general lifetime functions otherwise usually apply, and small-signal minority-carrier trap saturation obtains. The apparent diffusion-length lifetime then increases to a small-signal saturation-range value equal to the

majority-carrier release time.† Further increase occurs in the approach to a large-signal lifetime, which, in this case, is also the (small-signal) lifetime in the limit of strongly extrinsic material of the opposite conductivity type. Such increases of lifetime can account for certain cases of superlinearity, or the more-rapid-than-linear increase of photoconductivity with injection level, on the basis of a single trapping level.‡

Transient decay of photoconductivity is analyzed in Section 3.2. In the linear small-signal case, the decay is given by a sum of exponential modes with (real and positive) decay constants whose number exceeds by one the number of types of centers present.[2,28,37-41]§ For nonrecombinative trapping by centers of two types, the decay constants and equilibrium concentrations after injection are evaluated in Section 3.2.1 for electron and hole traps present together and for electron (or hole) traps only. With the latter, carriers released from one type may be captured in the other. The general linear case for centers of one type, including recombination, is analyzed in detail. The two time constants are given in forms involving the capture concentration. If one is large compared with the other, then the larger may be identified as the lifetime, while the smaller represents a trapping transient during which approach to the steady-state trapping ratio takes place. This transient has small amplitude for small concentration of centers, for which capture rates in the ratio of capture frequencies and release rates in the ratio of release frequencies decay with the concentration in the lifetime mode. It does not occur if the steady-state trapping ratio obtains initially, or if "critical recombination" obtains, with which, because of equal capture frequencies, trapped concentration does not change from initial value zero.

Sufficiently small capture concentration gives, with the comparatively short trapping transient, a lifetime substantially equal to[38,39,40] the common steady-state electron, hole and diffusion-length lifetime. The required condition is frequently severe: In the minority-carrier trapping range, it is the same as the common-lifetime condition. Capture concentration large results in decay times equal to[38,40] the steady-state electron and hole lifetimes and given by the electron and hole capture times. If one of these is large compared with the other, then the smaller represents the transient for practically complete trapping of the carriers of one kind, and the larger represents the recombinative decay of the

† Approximate steady-state solutions which exhibit small-signal nonlinearity have been given by Tolpygo and Rashba.[33]

‡ A multilevel model for superlinearity has been given by Rose,[29,34,35] (and Ref. 1, Ch. 1A). See also Bube.[35]

§ See also Ref. 1, Ch. 3A. This chapter also includes some nonlinear cases.

carriers in traps and of the carriers of the other kind as these are captured. In all these cases, the lifetime decreases monotonically as concentration of centers increases. Under a condition that is usually met in the minority-carrier trapping range, this decrease occurs primarily in two ranges of concentration of centers, with approximate constancy of lifetime in an intermediate range.†

The photoconductive decay is governed in the general case by nonlinear differential equations. These are considered for centers of a single type in Section 3.2.2. The general single-level problem is rather intractable analytically.‡ Solutions of the nonlinear equations are given for two special cases, namely, nonrecombinative trapping and sufficiently small concentration of centers or large concentrations of mobile excess carriers such that the steady-state lifetimes are substantially equal. The latter solution§ has the rather restricted general application of the common steady-state lifetime function,[31] since it is the integrated form corresponding to this function.‖ By solving suitably linearized equations, the decay times associated with a small-amplitude pulse of added carriers above a steady generation level are evaluated. If, as is often permissible, direct recombination may be neglected,[48,49,50] then the decay in the general large-signal limit is exponential with lifetime equal to the steady-state large-signal lifetime. During this decay, the concentrations of carriers in traps remain constant. This lifetime and the corresponding concentrations in traps are evaluated for centers of a single type and for the two-level cases. A differential equation that is invariant under interchange of quantities pertaining to electrons and to holes is derived for centers of a single type. It provides a first integral under a condition that holds for sufficiently large concentration of centers or concentrations of mobile carriers. With this first integral, the decay problem may be formulated as a first-order (rather than second-order) nonlinear differential equation. The large-signal condition, obtained in this connection, differs from the familiar one[10] in that, as a condition for equal electron and hole lifetimes, it entails not only relatively large change in conductivity but

† The approach to constancy with increasing concentration of centers is discussed by Wertheim.[41]

‡ Certain analytical approximations have been considered by Isay.[42] A treatment which includes numerically computed solutions has been given by Nomura and Blakemore.[43]

§ It is equivalent to ones given by Rittner, in Ref. 1, Ch. 3A, and by Guro.[44]

‖ The decay lifetime has been evaluated as this function by Okada.[45] That the nonlinearity according to this function does not account for (small-signal) decay in silicon has been observed by Blakemore.[46] This author has fitted dependences of lifetime on injection level and temperature assuming two-level recombination from one type of center or from two types. The common lifetime function has been employed for centers in germanium by Iglitsyn, Kontsevoi and Sidorov.[47] It appears that these centers were in the recombination range.

also requires saturation of centers, which may be present in relatively large concentration.

Conditions are obtained for centers that give recombination with substantially constant lifetime for minority carriers and inappreciable trapping. Constant lifetime that applies in the small-signal range also applies in the large-signal range, provided the energy level of the centers is not too far from the Fermi level towards the majority-carrier band. It requires, however, sufficient strongly extrinsic material. In material of mixed conductivity type, the recombination rate cannot, in general, be specified in terms of a minority-carrier lifetime. But "linear recombination" may apply, characterized by a two-lifetime recombination rate that is the sum of contributions respectively proportional to the added minority- and majority-carrier concentrations. The assumption of general linear recombination is also a convenient notational device: In the analysis of models involving nonrecombinative traps in conjunction with the recombination centers, it permits deriving results in forms that apply for any conductivity of either type.

The phenomenon of negative photoconductivity, or the decrease in conductivity below the equilibrium value upon optical injection,† results essentially from excitation of minority carriers from traps with recombination in other centers. Theory for this effect is given in Section 3.2.3, a general expression for mobile-carrier concentrations being derived for the linear small-signal case. This result is of comparative formal simplicity and shows that the effect tends to be offset by recombination in the traps and to be enhanced with deep traps of small capture cross section.‡

A general procedure is outlined in Section 3.2.4 for analysis of trapping models with a number of discrete energy levels, which relates the various decay times to capture cross sections and these energy levels. This procedure is applied to observations of Hornbeck and Haynes[54] on electron trapping in p-type silicon.§ For the sample on which the most extensive measurements were made, the decay times ranged from 20 microseconds to 260 seconds. Their model, that of two kinds of nonrecombinative traps with recombination in other centers, is found to imply a hole-capture cross section of the deep traps and of the shallower

---

† This has been analyzed by Stöckmann.[51] It has recently been observed in silicon by Collins.[52] Infrared quenching of photoconductivity or luminescence from short-wavelength excitation, discussed by Rose[29,34,35] and others, is a closely related effect.

‡ Excitations involving trapping levels may increase normal photoconductivity.[53]

§ See also Ref. 1, Ch. 3F.

(but still "deep") traps that is small compared with about 10^{-24} and 10^{-20} cm^2, respectively.

These cross sections are calculated from expressions for the decay times for nearly empty traps. Recombination in the deep traps cannot account for the observed decay: The hole-capture cross section that gives the decay time of 260 seconds for the nearly empty traps would give a considerably larger decay time, rather than the observed value of 1 second, for the traps nearly full. Recombination in the shallower traps, however, can account for the observed decay: The hole-capture cross section calculated from the decay time for these nearly empty traps is in close agreement with that which fits the entire decay in the deep traps. The lifetime with traps filled of 20 microseconds may then be ascribed to recombination in the higher level of two-level shallower traps. A recalculation of cross sections and energy levels on the basis of this model gives hole-capture cross sections large compared with 1.2×10^{-17} cm^2, equal to 2.4×10^{-20} cm^2 and small compared with 10^{-24} cm^2, respectively, for the recombination level and the shallower and deep trapping levels. The corresponding electron-capture cross sections† are 2.3×10^{-15}, 1.1×10^{-13} and 2.9×10^{-14} cm^2, the last two being half an order of magnitude smaller than the ones calculated by Hornbeck and Haynes. The shallower and deep trapping levels are found to lie 0.007 ev above and 0.23 ev below the Fermi level for intrinsic material; the latter trapping level is 0.78 ev below the conduction band.‡ Use is made of the observed straggle effect from the shallower traps, comparison being made with the theoretical expression derived in Section 3.4.3 for the limiting decay time at fixed location for the tail of the distribution from a pulse injected under applied field after the maximum has drifted past. It is shown that a model for which the trapping levels are levels of centers of a single type cannot account for the observations. While the levels found are close to two levels of gold,[56,57] it is thus unlikely that they result from a single metallic impurity. This conclusion bears on the indications that the deep traps are associated with the presence of oxygen as an impurity.

The steady-state photomagnetoelectric (PME) effect with trapping is analyzed in Section 3.3 on the basis of the general formulation with applied magnetic field.§ Equations formally similar to those for no trapping apply in terms of redefined quantities that involve the trapping

† Theory to account for such large cross sections has been given by Lax.[55]

‡ An energy gap of 1.10 ev at 300°K is used rather than 1.00 ev as in Ref. 54.

§ A treatment of photoconductance and PME voltage with trapping under ac illumination is included in: Lashkarev, Rashba, Romanov and Demidenko.[58] Mironov[59] deals with the transient decays after removal of steady illumination.

ratios. The effect may exhibit small-signal nonlinearity with nonuniform lifetime if recombinative deep traps in the minority-carrier trapping range are involved. The influence of trapping as such is investigated. It is found that nonrecombinative minority-carrier traps (in conjunction with recombination centers) increase diffusion-length lifetime by an amount proportional to the capture concentration.† Thus, minority-carrier trapping decreases PME current. A comparatively slight decrease in τ_0 and increase in PME current results from majority-carrier trapping.

Detailed illustrative procedures and related theory are given for the determination of capture cross sections, concentrations and energy levels from suitable PME and photoconductivity measurements at given temperature. With trapping in recombinative traps of a single type, the PME current-conductance ratio involves light intensity implicitly through its dependence on the lifetime τ_c that is defined in terms of the change in conductivity for a given steady, uniform volume-generation rate. The ratio, however, determines a relationship between τ_0 and τ_c, a transcendental relationship with the preferred method[11] of the high-recombination-velocity dark surface. This relationship, in conjunction with suitable additional conductance measurements also independent of light intensity, suffices to determine both τ_0 and τ_c, then the capture (and release) frequencies and capture concentration, and finally the quantities sought. The linear small-signal theory‡ is given for recombinative traps of a single type and also for nonrecombinative traps with recombination centers, for which the results are essentially similar though somewhat simpler.

Preliminary to analysis of transport problems, the general ambipolar continuity equation is specialized to the linear small-signal case in Section 3.4.1. Then, for trapping (and recombination) in centers of a

† Jonscher[60] gives an increase of diffusion length with trap concentration which is bounded and always essentially negligible, a result at variance with that given here. In Jonscher's nonambipolar treatment, the continuity equation does not include a term in the second space derivative of trapped-carrier concentration. Though this term is relatively small for sufficiently strongly extrinsic material, its neglect significantly affects the higher-order differential equation for concentration of mobile minority carriers, obtained by eliminating trapped-carrier concentration, in that it gives a coefficient of the term in the second space derivative that is too small by just the factor by which diffusion-length lifetime is increased.

‡ Zitter[61] discusses the phenomenological dependence for any model of electron and hole lifetimes on τ_c and lifetime derived from the PME effect (in the thick slab). The latter is the same as τ_0, and Zitter relates it to a diffusion length. Amith,[62,63] has presented the effect of nonrecombinative traps on the PME current-conductance ratio, and has pointed out that the predominant effect is usually on conductance. That on PME current is generally negligible in comparison if the traps are minority-carrier traps and are present in not too large concentration in sufficiently strongly extrinsic material.

single type and linear recombination in other centers, the respective concentrations are shown to satisfy certain third-order partial differential equations, which are of the second order in time. These reduce to the same equation if there is no volume generation. They otherwise each contain a term proportional to the volume generation function, the equations for the mobile-carrier concentrations containing the time derivative of this function as well.

The case analyzed in Section 3.4.2, that of injection into a filament in the steady state with applied field, yields qualitative information of interest. If a certain frequency ν_v, the "straggle constant," is positive, then the field-opposing and field-aiding solutions in the regions separated by the point of injection are sharply varying and gradually varying exponentials, as in the no-trapping case.[64] But with negative ν_v, gradually varying field-opposing and sharply varying field-aiding solutions obtain. In the limit of no diffusion, these give added carrier concentrations only in the direction opposite to the direction of drift normally determined by conductivity type. This "reverse drift" is explained by a simple and entirely general criterion, obtained from the fundamental equations, for the local direction of drift of a concentration disturbance: Normal or reverse drift occurs according to whether injection results in proportionately more or fewer minority carriers than there are at thermal equilibrium. For no trapping, for example, the concentrations are increased locally by the same increments, so that proportionately more minority carriers result if the material is extrinsic; and zero drift[10,64] obtains if the material is intrinsic. Conditions for the sign of ν_v are given. It is shown from these that reverse drift, which occurs for sufficiently large trap concentration in not too strongly extrinsic material, occurs with nonrecombinative trapping if minority carriers are trapped so that the fraction of the time they are free is smaller than the equilibrium minority-carrier to majority-carrier concentration ratio.

Drift of a pulse of carriers injected into a filament, with trapping by centers of a single type,† is analyzed in detail in Section 3.4.3. Bilateral

† Fan[37,65] has given a solution of this drift problem which applies for negligible majority-carrier capture frequency. Clarke[40] has, in effect, pointed out this restriction, to which solutions for the decay of photoconductivity given by Fan[37] and Rittner[1] are also subject. Jonscher[66] has given solutions for drift of minority carriers with recombination and nonrecombinative trapping at variance with solutions given here. The otherwise plausible neglect by Jonscher, in a nonambipolar treatment for strongly extrinsic material, of a term in the continuity equation involving the gradient of trapped-carrier concentration is apparently not justified. In the differential equation for concentration of mobile minority carriers, it results in minority-carrier release frequency only as a factor in the concentration-gradient term instead of ν_v, which, for this case, is substantially the sum of the capture and release frequencies. This neglect of the capture frequency is tantamount to neglect of the capture concentration compared to the equilibrium minor-

or two-sided Laplace transforms derived in Appendix A are used to obtain solutions of the differential equation for negligible diffusion. These solutions are of two main types, according to whether a frequency unit, ν, connected with the introduction of dimensionless variables and parameters, is real or imaginary. From theory concerning ν and the parameters, it is found that real ν implies either the minority-carrier trapping range or the recombination range.† Illustrative solutions of real ν for which nonrecombinative trapping is assumed are presented graphically. These show that carriers that remain untrapped appear in a comparatively rapidly attenuated pulse that drifts at the ambipolar velocity. This remnant of the initial pulse leads a continuous distribution, which, as a result of multiple trapping, ultimately spreads as a time-dependent gaussian distribution and exhibits a maximum that drifts at a fraction of the ambipolar velocity. For nonrecombinative trapping, this fraction approaches comparatively slowly a limiting value that does not exceed the fraction of the time the carriers are free,‡ and (for imaginary ν as well) the fraction of carriers trapped, obtained by integrating over the drift range, approaches comparatively rapidly the fraction of the time carriers are trapped. Recombination in other centers reduces the distance for a maximum at given time and thus the apparent mobility of the distribution. The decay constant for the straggle effect is found to be the (positive) straggle constant, accordingly so named.

Imaginary ν obtains over the majority-carrier trapping range and, for nonrecombinative trapping, over the reverse-drift range. With recombinative trapping, it obtains also for zero drift and over a normal-drift range other than that of majority-carrier trapping. Illustrative solutions, for which nonrecombinative trapping is assumed, are presented graphically for reverse drift and for majority-carrier trapping. It appears that in the reverse-drift range an attenuated pulse of untrapped carriers, which drifts at the ambipolar velocity, leads a continuous distribution

ity-carrier concentration. It presumably leads, for example, to the conclusion of this reference that a very short pulse is transmitted without distortion and is only attenuated. Also, the solution given for the steady state of continuous injection should properly include the fraction of the time minority carriers are free as a factor in the exponent. That is, trapping results in a more gradual decay with distance; for this case of no diffusion, a lifetime applies that is equal to the sum of the lifetime proper and the (generally much larger) lifetime for multiple trapping, which is discussed in Section 3.2.4.

† As shown in Section 2.2.1.2, these ranges in their entirety together constitute the "minority-carrier capture range," for which the equilibrium minority- to majority-carrier capture frequency ratio exceeds unity. It is shown in Section 3.4.3 that there is a minority-carrier capture range of imaginary ν which includes the reverse-drift range.

‡ Fan[37] has shown from his solution that this limiting value is, for relatively small trap concentration, equal to the free-time fraction.

of added mobile minority carriers, which crowds towards the injection point as its maximum excursions both above and below the axis increase with time. There is local carrier depletion, the distribution being negative over part of the drift range after a certain time. The distribution approaches a pulse at the injection point of strength equal, for nonrecombinative trapping, to the initial strength times the free-time fraction. It does not exhibit essentially unidirectional drift: The drift of added carriers, initially in the direction of the ambipolar velocity, is largely in the opposite direction after some trapping has taken place. A numerical estimate of the effect of diffusion indicates that negative added-carrier concentrations can occur over appreciable distances under conditions that can be realized in practice.† The illustrative solution for majority-carrier trapping shows that negative added-carrier concentrations occur in this case also. Majority-carrier trapping, however, results essentially in drift at the ambipolar velocity and, if it is nonrecombinative, the fraction of carriers trapped approaches the trapped-time fraction.

The solution is given for "critical trapping," the borderline case between cases of real and imaginary ν. For nonrecombinative trapping, it is the same as that for zero drift and gives exponential continuous distributions that are established progressively as the drift range increases and otherwise do not change with time. For trapping in intrinsic material without diffusion, drift does not start, and the initial pulse results simply in pulses for the concentration increments that remain at the injection point, where they change as trapping and recombination proceed. With diffusion, ambipolar drift occurs, since the condition for zero drift no longer holds in the intrinsic material as carriers are trapped away from the injection point. Further physical interpretations for the various cases are obtained by evaluating the current density of added carriers, which represents the equal departures for given total current density of the electron and hole flow densities from their values for no added carriers.

1.3 *List of Symbols*

The following list includes most of the symbols to be employed, and is largely consistent with previous notation.[10,11,64]

† Kaiser[67] has suggested that negative added-carrier concentrations that were observed with localized optical injection in silicon under applied field may be accounted for through these results. A theoretical discussion of carrier depletion is included in Ref. 10.

$a \equiv$ parameter in distribution of (153).

A_{nj}, $A_{pj} \equiv$ capture cross sections for electrons and holes for the jth energy level or type of center.

$b \equiv \mu_n/\mu_p$, drift mobility ratio.

$C_i \equiv$ coefficient for direct electron-hole recombination given in (37).

C_{nj}, $C_{pj} \equiv$ electron and hole capture coefficients for the jth energy level or type of center.

$D \equiv kT\mu_n\mu_p(n + p)/\sigma = (n + p)/(n/D_p + p/D_n)$, ambipolar diffusivity for no trapping.

$D' \equiv$ modified ambipolar diffusivity, defined in (31).

$D_i \equiv$ diffusivity for intrinsic material.

D_n, $D_p \equiv$ diffusion constants for electrons and holes.

$\hat{D}_n$, $\hat{D}_p \equiv$ diffusivities defined in (133).

$D_0 \equiv$ value of D at thermal equilibrium.

$D_0' \equiv$ value of D' at thermal equilibrium.

$e \equiv$ electronic charge.

$\mathbf{E} \equiv$ electrostatic field.

$\mathcal{E} \equiv$ Fermi level for intrinsic material.

$\mathcal{E}_F \equiv$ Fermi level.

$\mathcal{E}_F' \equiv$ "reflected Fermi level," the reflection of $\mathcal{E}_F$ about $\mathcal{E}$.

$\mathcal{E}_j \equiv$ electron energy for the jth energy level or type of center.

F_n, $\hat{F}_n$, $F_p \equiv$ fractions of mobile electrons, trapped electrons, and mobile holes for drift of a pulse, given by (164).

$g \equiv$ rate of volume generation of electron-hole pairs.

$\Delta g \equiv g - g_0$.

$g_0 \equiv$ value of g at thermal equilibrium.

$\mathbf{G} \equiv$ quantity defined by (9).

$\Delta G \equiv$ conductance increase of slab per unit width, given by (116), (118) and (129).

$G_0 \equiv$ dark conductance of slab per unit width.

$\mathbf{I} \equiv \mathbf{I}_p + \mathbf{I}_n$, total current density.

$\Delta\mathbf{I} \equiv$ current density of added carriers, defined by (19).

$\mathbf{I}_D \equiv$ diffusion current density, defined by (7).

$\mathbf{I}_n$, $\mathbf{I}_p \equiv$ electron and hole current densities.

I_0, $I_1 \equiv$ modified Bessel functions, in the notation of Watson.

J_0, $J_1 \equiv$ Bessel functions, in the notation of Watson.

$k \equiv$ Boltzmann's constant.

$\mathbf{k} \equiv$ unit vector in the direction of magnetic field.

$\mathbf{K} \equiv$ quantity defined by (10).
$K_G \equiv$ factor in (129) for conductance change of illuminated slab, evaluated in (130).
$K_\tau \equiv$ factor defined in (119) by which diffusion-length lifetime with nonrecombinative trapping exceeds recombination lifetime.
$L \equiv v_0\tau$, length unit defined in (143).
$\mathfrak{L} \equiv$ surface rate of generation of electron-hole pairs from strongly absorbed radiation.
$L_0 \equiv (D_0\tau_0)^{\frac{1}{2}}$, diffusion length.
$\mathfrak{L} \equiv$ operator symbol for two-sided Laplace transform.
$m \equiv p + \hat{p} = n + \hat{n}$.
$\Delta m \equiv m - m_0$.
$m_0 \equiv$ value of m at thermal equilibrium.
$n \equiv$ electron concentration.
$\Delta n \equiv n - n_0$.
$\overline{\Delta n} \equiv$ two-sided Laplace transform of Δn.
$\Delta N \equiv \Delta n/(\mathcal{P}/L)$.
$\overline{\Delta N} \equiv \overline{\Delta n}/(\mathcal{P}/L)$.
$\hat{n} \equiv$ concentration of fixed negative charges.
$\Delta\hat{n} \equiv \hat{n} - \hat{n}_0$.
$\overline{\Delta\hat{n}} \equiv$ two-sided Laplace transform of $\Delta\hat{n}$.
$\Delta\hat{N} \equiv \Delta\hat{n}/(\mathcal{P}/L)$.
$\mathfrak{N}, \mathfrak{N}_j \equiv$ total concentrations of centers.
$n_j \equiv$ electron concentration for the Fermi level at the jth trapping level.
$n_j^* \equiv$ equality density, defined in (54).
$\mathfrak{N}_j^* \equiv$ "capture concentration," defined (for $j = 1$) in (63).
$n_s \equiv n_0 - p_0$.
$n_0 \equiv$ value of n at thermal equilibrium.
$\hat{n}_0 \equiv$ value of $\hat{n}$ at thermal equilibrium.
$N_1, N_2 \equiv$ dimensionless decay constants defined by (147).
$p \equiv$ hole concentration.
$\Delta p \equiv p - p_0$.
$\overline{\Delta p} \equiv$ two-sided Laplace transform of Δp.
$\Delta P \equiv \Delta p/(\mathcal{P}/L)$.
$\overline{\Delta P} \equiv \overline{\Delta p}/(\mathcal{P}/L)$.
$\hat{p} \equiv$ concentration of fixed positive charges.
$\Delta\hat{p} \equiv \hat{p} - \hat{p}_0$.

$\mathcal{P} \equiv$ number per unit area of carrier pairs injected over cross section of filament.

$p_j \equiv$ hole concentration for the Fermi level at the jth trapping level.

$p_j^* \equiv$ equality density, defined in (54).

$p_0 \equiv$ value of p at thermal equilibrium.

$\hat{p}_0 \equiv$ value of $\hat{p}$ at thermal equilibrium.

$r_n \equiv d\hat{n}/dm$, steady-state trapping ratio.

r_{nj} , $r_{pj} \equiv$ trapping ratios for transient photoconductive decay modes, evaluated in Section 3.2.1; also similar quantities that are given in (149).

$r_p \equiv d\hat{p}/dm$, steady-state trapping ratio.

$\mathfrak{R}_m$, $\mathfrak{R}_n$, $\mathfrak{R}_p \equiv$ functions specifying rates of decrease of m, n and p through trapping and recombination.

$s \equiv$ Laplace transform variable.

s_m , s_n , $s_p \equiv$ surface recombination velocities for m, n and p, related by (115).

$T \equiv$ temperature in degrees absolute.

$U \equiv t/\tau$, dimensionless time variable.

$\mathbf{v} \equiv$ ambipolar drift velocity, defined by (8).

$\mathbf{v}' \equiv$ modified ambipolar drift velocity, defined in (31).

$\hat{\mathbf{v}}_n$, $\hat{\mathbf{v}}_p \equiv$ velocities defined in (133).

$\mathbf{v}_0 \equiv$ value of $\mathbf{v}$ at thermal equilibrium.

$\mathbf{v}_0' \equiv$ value of $\mathbf{v}'$ at thermal equilibrium.

$V \equiv$ electrostatic potential.

$X \equiv x/L$, dimensionless distance.

α_1 , $\alpha_2 \equiv$ quantities defined in (78).

$\alpha_{10} \equiv n_0/n_1 = p_1/p_0$.

$\alpha_{20} \equiv p_0/p_2 = n_2/n_0$.

$\Delta_1 \equiv$ quantity defined by (62).

$\zeta \equiv$ dimensionless parameter defined in (145).

$\eta \equiv$ dimensionless parameter defined by (152).

$\theta \equiv \theta_p - \theta_n = \theta_p + |\theta_n|$.

θ_n , $\theta_p \equiv$ Hall angles for electrons and holes.

$\Theta \equiv$ variable defined by (157).

$\kappa \equiv$ dimensionless parameter defined in (145).

μ_n , $\mu_p \equiv$ drift mobilities for electrons and holes.

$\nu \equiv$ frequency unit (real or imaginary) defined in (143).

$\nu_D \equiv$ quantity defined in (136).

ν_{gnj} , $\nu_{gpj} \equiv$ "effective" release frequencies for electrons and holes,

defined for acceptor-type ($j = 1$) and donor-type ($j = 2$) centers in (50), and for two-level centers in Section 2.2.2.

$\nu_{ij} \equiv$ decay constants defined by (47).

ν_{n3} , $\nu_{p3} \equiv$ decay constants for "linear recombination," defined in Section 3.2.2.

$\nu_s \equiv$ sum of hole and electron capture and release frequencies defined by (83).

ν_{tnj} , $\nu_{tpj} \equiv$ "effective" capture frequencies for electrons and holes, defined for acceptor-type ($j = 1$) and donor-type ($j = 2$) centers in (50), and for two-level centers in Section 2.2.2.

$\nu_v \equiv$ "straggle constant," defined in (136).

ν_1 , $\nu_2 \equiv$ decay constants for photoconductivity.

$\xi \equiv$ dimensionless parameter defined by (151).

$\sigma \equiv \sigma_n + \sigma_p$, total conductivity.

σ_n , $\sigma_p \equiv$ partial conductivities for electrons and holes.

$\tau \equiv$ time unit defined in (143).

$\tau_c \equiv$ conductivity lifetime, defined by (117).

τ_{gnj} , $\tau_{gpj} \equiv$ "effective" release times for electrons and holes, the reciprocals of ν_{gnj} , ν_{gpj} .

$\tau_m \equiv$ steady-state lifetime or lifetime function for Δm.

$\bar{\tau}_m \equiv$ lifetime for Δm with recombination centers as well as traps, evaluated in Section 3.3.

τ_n , $\tau_p \equiv$ steady-state electron and hole lifetimes or lifetime functions.

τ_{n0} , $\tau_{p0} \equiv (C_{nj}\mathfrak{N}_j)^{-1}$, $(C_{pj}\mathfrak{N}_j)^{-1}$ for a particular j.

$\tau_r \equiv$ apparent lifetime from PME current-conductance ratio, evaluated in Section 3.3.

τ_{tnj} , $\tau_{tpj} \equiv$ "effective" capture times for electrons and holes, the reciprocals of ν_{tnj} , ν_{tpj} .

$\tau_0 \equiv$ "diffusion-length lifetime," evaluated in (35) and (65).

$\bar{\tau}_0 \equiv$ "diffusion-length lifetime" with nonrecombinative traps and recombination centers, given by (119).

τ_1 , $\tau_2 \equiv$ time constants for photoconductive decay, the reciprocals of ν_1 , ν_2 .

$\tau_3 \equiv$ lifetime for decay through recombination centers, the reciprocal of ν_{n3} or ν_{p3} .

$\tau_\infty \equiv$ photoconductive decay time for nearly empty traps.

$\Phi \equiv -e^{-1}\mathcal{E}$.

$\Psi \equiv$ potential defined by (15).

$\Psi_j \equiv -e^{-1}\mathcal{E}_j$.

II. GENERAL FORMULATION

2.1 *The Transport Equations*

The general neutrality condition may be written as

$$m \equiv p + \hat{p} = n + \hat{n}, \tag{1}$$

which states that the total concentration of positive charges, the sum of the concentrations of mobile holes and fixed positive charges, is equal to the corresponding total concentration of negative charges. It is, as will appear, of advantage to deal with the total concentration m of charges of either kind.

By way of extension of the familiar (nonambipolar) continuity equations for holes and for electrons that apply for no trapping, two forms of the continuity equation for m may be written:

$$\begin{aligned} \partial m/\partial t &= \partial p/\partial t + \partial \hat{p}/\partial t = -e^{-1} \operatorname{div} \mathbf{I}_p + g - \mathfrak{R}_m \\ &= \partial n/\partial t + \partial \hat{n}/\partial t = e^{-1} \operatorname{div} \mathbf{I}_n + g - \mathfrak{R}_m . \end{aligned} \tag{2}$$

Here, for simplicity, the same volume-generation-rate function g is assumed for both holes and electrons; generalization to include excitation to or from trapping levels (as well as interband excitation) is given in Section 2.2.3. The volume rate $\mathfrak{R}_m$ is associated with trapping and recombination. It depends directly on the various concentrations and not explicitly on coordinates and time; $\partial \hat{p}/\partial t$ and $\partial \hat{n}/\partial t$ contribute only to $\mathfrak{R}_m$, and, if these are respectively subtracted from (2), then continuity equations for holes and electrons, namely,

$$\begin{aligned} \partial p/\partial t &= -e^{-1} \operatorname{div} \mathbf{I}_p + g - \mathfrak{R}_p , \qquad & \mathfrak{R}_p &= \mathfrak{R}_m + \partial \hat{p}/\partial t, \\ \partial n/\partial t &= e^{-1} \operatorname{div} \mathbf{I}_n + g - \mathfrak{R}_n , \qquad & \mathfrak{R}_n &= \mathfrak{R}_m + \partial \hat{n}/\partial t \end{aligned} \tag{3}$$

result. The same volume rate $\mathfrak{R}_m$ is properly used in each of the equations (2) since it depends directly only on concentrations; it must apply, in particular, in the case of zero current densities. As (2) shows, this use of $\mathfrak{R}_m$ is consistent with the neutrality condition and with the condition[64]

$$\operatorname{div} \mathbf{I} = 0, \qquad \mathbf{I} = \mathbf{I}_p + \mathbf{I}_n , \tag{4}$$

which applies in regions containing no sources or sinks of current. Differing volume rates for $p + \hat{p}$ and $n + \hat{n}$ are properly introduced only if there is appreciable space charge.

The familiar current-density equations,

$$\mathbf{I}_p = \sigma_p \mathbf{E} - eD_p \operatorname{grad} p$$

and (5)

$$\mathbf{I}_n = \sigma_n \mathbf{E} + eD_n \operatorname{grad} n,$$

apply under the assumption of Boltzmann statistics, which imply also proportionality of the hole and electron mobilities μ_p and μ_n to the corresponding diffusion constants D_p and D_n in accordance with Einstein's relation. By use of these equations and the neutrality condition, (1), a continuity equation for m of ambipolar form may be derived: The hole and electron current densities $\mathbf{I}_p$ and $\mathbf{I}_n$ are eliminated from (2); the electrostatic field $\mathbf{E}$ is also eliminated by means of the expression for $\mathbf{E}$ involving the total current density $\mathbf{I}$ that is obtained by adding the equations in (5); and use is made of (4), the condition of solenoidal $\mathbf{I}$. This procedure is similar to that previously employed in the no-trapping case[10] except that, for the required generality, p and n are treated formally as unrelated variables. The single continuity equation for m that results from (2) may be written in various forms as follows:†

$$\begin{aligned} \partial m/\partial t - g + \mathfrak{R}_m &= -e^{-1} \operatorname{div} \mathbf{I}_D - \mathbf{v}\cdot\mathbf{G} \\ &= -e^{-1}(\operatorname{div} \mathbf{I}_D + \mathbf{K}\cdot\mathbf{I}) \\ &= -e^{-1} \operatorname{div} \mathbf{I}^*, \end{aligned} \tag{6}$$

where

$$\begin{aligned} \mathbf{I}_D &\equiv -e\sigma^{-1}(\sigma_p D_n \operatorname{grad} n + \sigma_n D_p \operatorname{grad} p) \\ &= -e[D \operatorname{grad} m - \sigma^{-1}(\sigma_p D_n \operatorname{grad} \hat{n} + \sigma_n D_p \operatorname{grad} \hat{p})] \\ &= -eD(n+p)^{-1} \operatorname{grad} np = -ekT\mu_n\mu_p\sigma^{-1} \operatorname{grad} np, \end{aligned} \tag{7}$$

$$\mathbf{v} \equiv e\mu_n\mu_p(n-p)\sigma^{-2}\mathbf{I}, \tag{8}$$

$$\begin{aligned} \mathbf{G} &\equiv (n-p)^{-1}(n \operatorname{grad} p - p \operatorname{grad} n) \\ &= \operatorname{grad} m - (n-p)^{-1}(n \operatorname{grad} \hat{p} - p \operatorname{grad} \hat{n}), \end{aligned} \tag{9}$$

$$\begin{aligned} \mathbf{K} &\equiv (D_n - D_p)^{-1} \operatorname{grad} D = e^2\mu_n\mu_p\sigma^{-2}(n \operatorname{grad} p - p \operatorname{grad} n) \\ &= e(\mu_p^{-1} - \mu_n^{-1})^{-1} \operatorname{grad} [\sigma^{-1}(n+p)] \\ &= -e(\mu_p^{-1} + \mu_n^{-1})^{-1} \operatorname{grad} [\sigma^{-1}(n-p)], \end{aligned} \tag{10}$$

and

$$\mathbf{I}^* \equiv \alpha\mathbf{I}' + \beta\mathbf{I}'', \qquad \alpha + \beta = 1, \tag{11}$$

† This equation specialized to the case of $\Delta\hat{p} = 0$ can be shown to be consistent with a continuity equation for Δp derived by Rittner[1] under the assumption of a common lifetime function for electrons and holes.

with

$$\begin{aligned}\mathbf{I}' &\equiv \mathbf{I}_D - e(\mu_p^{-1} + \mu_n^{-1})^{-1}\sigma^{-1}(n-p)\mathbf{I} \\ &= -e(\mu_p^{-1} + \mu_n^{-1})^{-1}[(n-p)\mathbf{E} + (kT/e)\ \text{grad}\ (n+p)] \\ &= (b+1)^{-1}(b\mathbf{I}_p - \mathbf{I}_n) \qquad (12) \\ \mathbf{I}'' &\equiv \mathbf{I}_D + e(\mu_p^{-1} - \mu_n^{-1})^{-1}\sigma^{-1}(n+p)\mathbf{I} \\ &= e(\mu_p^{-1} - \mu_n^{-1})^{-1}[(n+p)\mathbf{E} + (kT/e)\ \text{grad}\ (n-p)] \\ &= (b-1)^{-1}(b\mathbf{I}_p + \mathbf{I}_n).\end{aligned}$$

The diffusivity D, a definition of which is contained in the last of the equations in (7), is the general ambipolar concentration-dependent diffusivity, which occurs in the theory for no trapping. It is used here simply for notational convenience. The velocity $\mathbf{v}$ of (8) is properly interpreted in the continuity equation as two velocities, $n\mathbf{v}/(n-p)$ for drift of Δp and $p\mathbf{v}/(p-n)$ for drift of Δn. It is otherwise formally similar to the ambipolar velocity of the theory for no trapping, except that now $n - p$ is not a constant concentration. The first two right-hand forms of the continuity equation, (6), exhibit terms associated with diffusion and drift, respectively, as comparison with the continuity equation for the no-trapping case shows. The current density $\mathbf{I}^*$, as given by (11), is introduced for generality; with solenoidal $\mathbf{I}$, the divergences of $\mathbf{I}'$ and $\mathbf{I}''$ are equal.† From the expressions for these current densities in terms of $\mathbf{I}_p$, $\mathbf{I}_n$ and the drift mobility ratio b given in (12), it may be verified that $\mathbf{I}^*$ may be chosen as $\mathbf{I}_p$ or $-\mathbf{I}_n$, as in (2). Indeed, as is otherwise evident, $\mathbf{I}^*$ may be written simply as a linear combination of $\mathbf{I}_p$ and $-\mathbf{I}_n$, normalized as in (11), since a linear combination, so normalized, of any two $\mathbf{I}^*$ is also an $\mathbf{I}^*$. The current densities $\mathbf{I}'$ and $\mathbf{I}''$ are introduced because their use is frequently convenient.

The mobile-carrier concentrations n and p are, in accordance with (1), properly written as $m - \hat{n}$ and $m - \hat{p}$ where they occur explicitly and in the diffusivity D, in the electron and hole conductivities $\sigma_n = e\mu_n n$ and $\sigma_p = e\mu_p p$ and in the total conductivity $\sigma = \sigma_n + \sigma_p$. For acceptor and donor centers of single types, $\partial\hat{n}/\partial t$ and $\partial\hat{p}/\partial t$ in terms of the various concentrations provide, with the continuity equation, three simultaneous differential equations in the dependent variable m, $\hat{n}$ and $\hat{p}$. For more than single types of acceptor and donor centers, $\hat{n}$ and $\hat{p}$ are sums of fixed-charge concentrations. Equations are then written for the rates of increase of each of these concentrations, and the number of

† Note that these divergences equal that of $\mathbf{I}_p$ and $(-\mathbf{I}_n)$; also, $\mathbf{I}'' - \mathbf{I}'$ equals $2b(b^2-1)^{-1}\mathbf{I}$.

simultaneous differential equations exceeds by one the total number of types of centers present. These are the differential equations for the general transport problem with trapping and recombination provided **I** is a known function of the space coordinates and time.

In some cases, **I** must be determined from boundary conditions. Use is then made of the fundamental differential equation, (4), which may be written to involve the electrostatic potential V as additional dependent variable. With

$$\sigma\mathbf{E} = -\sigma \text{ grad } V = \mathbf{I} - e \text{ grad } (D_n n - D_p p), \tag{13}$$

it follows that (4) may be written in the form

$$\text{div } [\sigma \text{ grad } V - (kT/e) \text{ grad } (\sigma_n - \sigma_p)] = 0, \tag{14}$$

in which σ_n , σ_p and σ are to be expressed in terms of m, $\hat{n}$ and $\hat{p}$. In this formulation, V is introduced into the continuity equation through the elimination of **I** by means of (13). Another procedure, of advantage in some connections, involves use of the potential

$$\Psi \equiv V - (kT/e)(b - 1)(b + 1)^{-1} \ln (\sigma/\sigma_0) \tag{15}$$

instead of V as dependent variable. Then **I** is given by

$$\mathbf{I} = -\sigma \text{ grad } \Psi + eD_i \text{ grad}(\hat{p} - \hat{n}), \tag{16}$$

where $D_i \equiv 2(D_p^{-1} + D_n^{-1})^{-1}$ is the diffusivity in intrinsic material. Aside from the effect of trapping on **I**, as given by the second term of (16), Ψ is the potential that "drives" the total current density. This may be described as the electrostatic potential modified by the Dember potential. The latter gives the field associated with diffusion of carriers of differing diffusion constants.

Electrostatic field given by

$$\mathbf{E} = e^{-1}(n + p)^{-1}[I_n/\mu_n + \mathbf{I}_p/\mu_p - kT \text{ grad}(\hat{p} - \hat{n})], \tag{17}$$

an equation somewhat analogous to (16), is a result obtained by solving for **E** in the equations for $\mathbf{I}''$ in (12). As (17) shows, **E** in the absence of trapping (and of appreciable space charge) may be written in a form that does not involve concentration gradients explicitly.

In the ambipolar form of the present treatment, the equations of (5) are

$$\mathbf{I}_p = (\sigma_p/\sigma)\mathbf{I} + \mathbf{I}_D = (\sigma_{p0}/\sigma_0)\mathbf{I} + \Delta\mathbf{I}$$

and (18)

$$\mathbf{I}_n = (\sigma_n/\sigma)\mathbf{I} - \mathbf{I}_D = (\sigma_{n0}/\sigma_0)\mathbf{I} - \Delta\mathbf{I},$$

in which zero subscripts denote values at thermal equilibrium; $\Delta\mathbf{I}$ is defined by

$$\begin{aligned}\Delta\mathbf{I} &\equiv e^2\mu_n\mu_p\sigma_0^{-1}\sigma^{-1}(n_0p - p_0n)\mathbf{I} + \mathbf{I}_D \\ &= \tfrac{1}{2}e\sigma_0^{-1}[(\mu_n + \mu_p)(n_0 + p_0)\mathbf{I}' + (\mu_n - \mu_p)(n_0 - p_0)\mathbf{I}''] \\ &= \sigma_0^{-1}(\sigma_{n0}\mathbf{I}_p - \sigma_{p0}\mathbf{I}_n). \qquad (19)\end{aligned}$$

Application of one or more of (13) or (17) and (18) is frequently required in connection with boundary conditions. The ambipolar diffusion current density $\mathbf{I}_D$ includes the effect of the Dember field and contributes the same particle flow density to both $\mathbf{I}_n$ and $\mathbf{I}_p$. Use of the expressions for $\mathbf{I}_n$ and $\mathbf{I}_p$ that involve $\Delta\mathbf{I}$ is of particular advantage for physical interpretations and in small-signal cases, since $\Delta\mathbf{I}$ is the current density of excess mobile carriers.[10] For given total current density $\mathbf{I}$, it represents the equal electron and hole flow densities, that are the departures from the thermal-equilibrium flow densities and that do not contribute to $\mathbf{I}$. Note that $\mathbf{I}^*$ may also be chosen as $\Delta\mathbf{I}$.

2.1.1 *Extension for Applied Magnetic Field*

The current densities for Hall angles θ_n and θ_p small are given in general by Equations (10) and (13) of a previous paper.[11] These result in

$$\begin{aligned}e(\partial m/\partial t - g + \mathfrak{R}_m) &= -\operatorname{div}\mathbf{I}_p = \operatorname{div}\mathbf{I}_n \\ &= -\operatorname{div}(\sigma_p\mathbf{E}) + eD_p \operatorname{div}\operatorname{grad} p \\ &\qquad - \theta_p[\![\operatorname{grad}\sigma_p, \mathbf{E}, \mathbf{k}]\!] \\ &= \operatorname{div}(\sigma_n\mathbf{E}) + eD_n \operatorname{div}\operatorname{grad} n \\ &\qquad + \theta_n[\![\operatorname{grad}\sigma_n, \mathbf{E}, \mathbf{k}]\!], \qquad (20)\end{aligned}$$

in which $\mathbf{k}$ is a unit vector in the direction of the magnetic field and the heavy brackets denote scalar triple products. With n and p treated formally as unrelated variables, multiplying respectively by σ_n and σ_p, adding and simplifying gives

$$\begin{aligned}\partial m/\partial t - g + \mathfrak{R}_m &= -e^{-1}(\operatorname{div}\mathbf{I}_D + \mathbf{K}\cdot\mathbf{I}) \\ &\quad - e^2\mu_n\mu_p\sigma^{-3}[\theta[\![(\mu_p p^2 \operatorname{grad} n + \mu_n n^2 \operatorname{grad} p), \mathbf{I}, \mathbf{k}]\!] \\ &\quad + (\theta_p\sigma_n + \theta_n\sigma_p)(D_n n - D_p p)[\![\operatorname{grad} n, \operatorname{grad} p, \mathbf{k}]\!]], \qquad (21)\end{aligned}$$

where $\mathbf{I}_D$ and $\mathbf{K}$ are defined in (7) and (10) and θ is the sum of the magnitudes of the Hall angles, $\theta_p - \theta_n$. In deriving this continuity

equation, use is made of curl $\mathbf{E} = 0$, which holds for steady applied magnetic field; time dependence of $\mathbf{I}$ generally has a quite negligible effect.[11] Use is also made of the relationships

$$\sigma\mathbf{E} = \mathbf{I} - e\,\text{grad}(D_n n - D_p p) - \sigma^{-1}(\theta_p\sigma_p + \theta_n\sigma_n)\mathbf{I} \times \mathbf{k} - \theta\mathbf{I}_D \times \mathbf{k} \quad (22)$$

and

$$[\![(\theta_p n\,\text{grad}\,p - \theta_n p\,\text{grad}\,n),\mathbf{E},\mathbf{k}]\!] = \sigma^{-1}[\![(\theta_p n\,\text{grad}\,p - \theta_n p\,\text{grad}\,n),\mathbf{I},\mathbf{k}]\!] + e\sigma^{-1}(\theta_p D_n n - \theta_n D_p p)[\![\text{grad}\,n,\,\text{grad}\,p,\mathbf{k}]\!], \quad (23)$$

which hold with the neglect of terms quadratic in Hall angles. Equation (23) is obtained in a straightforward manner from (22), which is obtained by writing total current density as

$$\mathbf{I} = \sigma\mathbf{E} + e\,\text{grad}(D_n n - D_p p) + (\theta_p\sigma_p + \theta_n\sigma_n)\mathbf{E} \times \mathbf{k} - e\,\text{grad}(\theta_p D_p p - \theta_n D_n n) \times \mathbf{k} \quad (24)$$

and then solving for $\mathbf{E}$. The terms on the right-hand side of (24) represent, respectively, drift, Dember, Hall and PME contributions.

A differential equation that expresses the solenoidal property of $\mathbf{I}$ is, from (24),

$$\text{div}[\sigma\mathbf{E} + (kT/e)\,\text{grad}(\sigma_n - \sigma_r)] + [\![\text{grad}(\theta_p\sigma_p + \theta_n\sigma_n),\mathbf{E},\mathbf{k}]\!] = 0. \quad (25)$$

If direct use must be made of this fundamental equation, then it is well to eliminate $\mathbf{I}$ from (21) by means of (24), and to employ the electrostatic potential V as one of the dependent variables.

The current densities are given in ambipolar form by

$$\mathbf{I}_p = (\sigma_p/\sigma)\mathbf{I} + \mathbf{I}^{\equiv}$$

and (26)

$$\mathbf{I}_n = (\sigma_n/\sigma)\mathbf{I} - \mathbf{I}^{\equiv},$$

where, if terms quadratic in Hall angles are neglected,

$$\mathbf{I}^{\equiv} \equiv \mathbf{I}_D + \theta(\sigma_n\sigma_p/\sigma^2)\mathbf{I} \times \mathbf{k} - \sigma^{-1}(\theta_p\sigma_n + \theta_n\sigma_p)\mathbf{I}_D \times \mathbf{k}. \quad (27)$$

Components of total current density perpendicular to the applied magnetic field are

$$I_x = \sigma E_x + e\frac{\partial}{\partial x}(D_n n - D_p p) + \sigma^{-1}(\theta_p\sigma_p + \theta_n\sigma_n)I_y + \theta I_{Dy} \quad (28)$$

and I_y, which is given by a similar expression obtained by interchanging x's and y's and (to retain a right-handed coordinate system by effectively reversing the direction of the z axis) changing the signs of the Hall angles. One way of deriving (28) is to substitute the expression obtained by solving for E_y in the equation for I_y, for E_y in the equation for I_x obtained from (24), and to neglect terms quadratic in Hall angles.

2.1.2 *Formulation for the Steady State in Terms of Trapping Ratios*

A number of results for the steady state can be established from the general differential equations without specifying in detail the trapping and recombination statistics. Differential "trapping ratios"

$$r_n \equiv d\hat{n}/dm, \qquad r_p \equiv d\hat{p}/dm \quad (29)$$

are introduced. These apply since, in the steady state, $\hat{n}$ and $\hat{p}$ each depend directly only on total concentration m of negative or positive charges. In the immediate context, r_n and r_p will be considered simply as factors that depend in general on m, which, multiplying grad m, give grad $\hat{n}$ and grad $\hat{p}$, respectively. They apply, of course, for any number of types of centers present. Their evaluation for particular models is given in Section 3.1 in connection with the more detailed analysis of the steady state.

With (29), it follows from (6) through (9) that the continuity equation for the steady state may be written as

$$\text{div}\,(D'\,\text{grad}\,\Delta m) - \mathbf{v}'\cdot\text{grad}\,\Delta m + \Delta g - \Delta m/\tau_m = 0, \quad (30)$$

in which D' and $\mathbf{v}'$ are modified ambipolar diffusivity and drift velocity that are given by

$$\begin{aligned} D' &\equiv kT\mu_n\mu_p\sigma^{-1}[(1 - r_p)n + (1 - r_n)p] \\ &= [1 - (r_p n + r_n p)/(n + p)]D, \\ \mathbf{v}' &\equiv e\mu_n\mu_p\sigma^{-2}[(1 - r_p)n - (1 - r_n)p]\mathbf{I} \\ &= [1 - (r_p n - r_n p)/(n - p)]\mathbf{v}, \end{aligned} \quad (31)$$

and in which the net generation rate $g - \mathcal{R}_m$ has been written as the increment in this rate over thermal equilibrium, $\Delta g - \Delta m/\tau_m$, with Δg and Δm being the corresponding increments in g and m and τ_m a lifetime function for Δm. The modified diffusivity and velocity do not

apply to time-dependent cases; $\mathbf{v}'$ would, for example, give the effect of applied field on apparent diffusion length, but is not, as will appear in Section 3.4.3, drift velocity for an injected pulse.

Expressing the concentration gradients for the steady state in terms of r_n or r_p and grad Δm also formally simplifies (14), the differential equation that must also be used if the current-flow geometry is not known. In connection with (18), the current-density equations, the procedure results in ambipolar diffusion current density given by

$$\mathbf{I}_D = -eD' \text{ grad } \Delta m \tag{32}$$

for the steady state.

The trapping ratios defined by (29) can assume negative as well as positive values: If centers of a given type trap mostly carriers of the opposite charge, then a negative trapping ratio obtains. Consider, for example, trapping in centers of the acceptor type, which are neutral or negatively charged. For these, positive r_n cannot exceed unity; it nearly equals unity if electron trapping is the predominant process, so that the excess trapped electron and mobile hole concentrations are substantially equal. If, however, hole trapping is the predominant process, then r_n is a large negative number, the increment in concentration of fixed negative charges being negative and balanced by the excess mobile electron concentration, so that m retains substantially its thermal-equilibrium value. Similar considerations apply to r_p for centers of the donor type. Thus, the trapping ratio is close to unity or a large negative number according to whether the centers predominantly trap carriers of the same charge or of the opposite charge.†

For a large negative trapping ratio, the comparatively small increments in m are associated with large magnitudes of D' and $\mathbf{v}'$, as (31) shows. A concentration variable other than Δm may then be more suitable. The equation in the linear combination $A\Delta n + B\Delta p$ (with constant A and B) of the excess mobile-carrier concentrations that results from (30) has diffusivity and velocity equal to D' and $\mathbf{v}'$ each divided by $A(1 - r_n) + B(1 - r_p)$, since, from (29), grad Δm is $[A(1 - r_n) + B(1 - r_p)]^{-1}$ grad $(A\Delta n + B\Delta p)$. In general, they are bounded in magnitude for all values of the trapping ratios that can occur.‡ In this equa-

† For acceptor centers, say, of total concentration $\mathfrak{N}_1$, the trapping ratio $r_p' \equiv d(\mathfrak{N}_1 - \hat{n})/d(p + \mathfrak{N}_1 - \hat{n})$ for holes may be defined. The two ratios are symmetrically related: They may be interchanged in $r_p' = r_n/(r_n - 1)$ and, as one increases to unity, the other becomes negatively infinite.

‡ Note that $A(1 - r_n) + B(1 - r_p)$ equals $r_p - r_n$ for $A = 1$ and $B = -1$ and, if r_n and r_p are constants (as obtains under suitable small-signal restriction), also for $A = r_p$ and $B = -r_n$. It follows that $n - p$ and $r_p n - r_n p$ are (under this restriction) both subject to diffusivity $D_0'/(r_p - r_n)$ and velocity $\mathbf{v}_0'/(r_p - r_n)$.

tion, the recombination term is properly written as $-(A\Delta n + B\Delta p)/(A\tau_n + B\tau_p)$, where τ_n and τ_p are lifetime functions for Δn and Δp. These functions are respectively equal to $\Delta n/\mathcal{R}_m$ and $\Delta p/\mathcal{R}_m$ since, from (3), $\mathcal{R}_n = \mathcal{R}_p = \mathcal{R}_m$ holds for the steady state.

The equilibrium lifetimes for electrons and holes differ in general, but are nevertheless always associated with the same diffusion length. This result follows readily from (30), whose linear small-signal form is

$$D_0' \text{ div grad } \Delta m - \mathbf{v}_0' \cdot \text{grad } \Delta m + \Delta g - \Delta m/\tau_m = 0, \qquad (33)$$

the zero subscripts denoting thermal-equilibrium values.† The lifetime function τ_m is here constant; and, since Δn and Δp equal $(1 - r_n)\Delta m$ and $(1 - r_p)\Delta m$, with r_n and r_p the thermal-equilibrium trapping ratios, (33) implies

$$\begin{aligned}(1 - r_n)^{-1} D_0' \text{ div grad } \Delta n - (1 - r_n)^{-1}\mathbf{v}_0' \cdot \text{grad } \Delta n \\ + \Delta g - \Delta n/(1 - r_n)\tau_m = 0\end{aligned} \qquad (34)$$

for electrons and a similar equation for holes. Thus, for Δn the lifetime is τ_m multiplied by $(1 - r_n)$, while — as may be established in greater generality from (29) and (30) — the diffusivity and velocity are those for Δm multiplied by the reciprocal of this factor, and similarly for Δp. It follows, in particular, that the product of equilibrium diffusivity and lifetime, which is the square of L_0, the diffusion length, is the same for Δn, Δp and Δm, independently of the particular trapping and recombination statistics.[9] A "diffusion-length lifetime" τ_0, based on the unmodified ambipolar diffusivity D_0, may accordingly be defined by[61]

$$\begin{aligned}\tau_0 \equiv L_0^2/D_0 = (D_0'/D_0)\tau_m &= [1 - (r_p n_0 + r_n p_0)/(n_0 + p_0)]\tau_m \\ &= (n_0\tau_p + p_0\tau_n)/(n_0 + p_0),\end{aligned} \qquad (35)$$

in which τ_p and τ_n are the equilibrium lifetimes for Δp and Δn. The more detailed analysis of Section 3.1.1 includes evaluation of the single diffusion length and lifetime τ_0 that correspond to the (equilibrium) Shockley-Read electron and hole lifetimes. Diffusion-length lifetime for recombination in the presence of nonrecombinative traps is evaluated in Section 3.3.

For the steady-state formulation that includes applied magnetic field, it is readily shown that (21), the continuity equation, assumes the form of (30) if $\mathbf{v}'$ is redefined in accordance with

† As shown in Section 3.1.1, the required small-signal restriction may be more severe than that given in Ref. 10 for the no-trapping case.

$$\mathbf{v}' \equiv e\mu_n\mu_p\sigma^{-2}\{[(1 - r_p)n - (1 - r_n)p]\mathbf{I} + \theta(e/\sigma)[(1 - r_p)\mu_n n^2 + (1 - r_n)\mu_p p^2]\mathbf{I} \times \mathbf{k}\}. \tag{36}$$

Note that the second scalar product in (21) vanishes, since steady-state concentration gradients are collinear vectors. For this case, use of the trapping ratios formally simplifies (25), as well as (27) and (28), which involve the form for $\mathbf{I}_D$ of (32).

2.2 *Mass-Action Theory*

2.2.1 *Single-Level Centers of Two Types*

In this section, centers of both the acceptor and donor types are assumed to be present, namely centers that can have respectively single negative or positive charges or be neutral. By use of a suitable convention, the equations apply, in effect, to the more general model of two types of centers each of which has two states of charge (which differ by one electronic charge). On the basis of equations of this section, theory for centers of a single type but with two energy levels or three states of charge is given in Section 2.2.2.

Under the assumption of mass-action interactions, the equations

$$\begin{aligned} g - \mathfrak{R}_m &= g - C_i np - C_{p1}[p\hat{n} - p_1(\mathfrak{N}_1 - \hat{n})] \\ &\quad - C_{n2}[n\hat{p} - n_2(\mathfrak{N}_2 - \hat{p})], \\ \partial\hat{n}/\partial t = \mathfrak{R}_n - \mathfrak{R}_m &= C_{n1}[n(\mathfrak{N}_1 - \hat{n}) - n_1\hat{n}] \\ &\quad - C_{p1}[p\hat{n} - p_1(\mathfrak{N}_1 - \hat{n})], \\ \partial\hat{p}/\partial t = \mathfrak{R}_p - \mathfrak{R}_m &= -C_{n2}[n\hat{p} - n_2(\mathfrak{N}_2 - \hat{p})] \\ &\quad + C_{p2}[p(\mathfrak{N}_2 - \hat{p}) - p_2\hat{p}] \end{aligned} \tag{37}$$

hold. The first equation gives $\mathfrak{R}_m$, and it (as well as the other two) is obtained by considering the photoconductive case of uniform concentration and no transport, $g - \mathfrak{R}_m$ being the contribution to $\partial m/\partial t$ that does not involve transport. Four processes are taken into account for each type of center. In the second equation, for example, the term $C_{p1}p\hat{n}$ is the volume rate of neutralization of fixed negative charges by holes; C_{p1} is a phenomenological capture coefficient,† which depends in general on temperature and not on concentration. The second term in the same brackets gives the rate for the inverse process, $C_{p1}p_1$ being the emission

† In the terminology of Sah and Shockley[22] this quantity is called a capture probability.

coefficient for hole emission from a neutral acceptor center. Here $\mathfrak{N}_1$ is the total concentration of the acceptor centers, and the concentration p_1, constant at given temperature, is defined by the condition that the quantity in brackets vanishes at thermal equilibrium, in accordance with detailed balance. The preceding brackets relate to the interactions of the same centers with electrons, the term $C_{n1}n(\mathfrak{N}_1 - \hat{n})$ being the volume rate of capture of electrons by the neutral acceptor centers and $C_{n1}n_1$ being the coefficient for electron emission from the charged ones. The concentrations n_1 and p_1 are those of the Hall-Shockley-Read theory, and are here introduced without explicit reference to Boltzmann statistics.[2,13,14] The third equation expresses the dependence of $\partial\hat{p}/\partial t$ on the analogous processes for the donor centers. In the first equation, which includes the rate $C_i np$ of direct electron-hole recombination, only interactions that change the total concentration m are involved.

The sign or magnitude of the charge that a center can assume is not of material significance in the analysis of this section; although written symmetrically for fixed charges of both signs, (37) may formally be transformed so as to apply to two types of donor or acceptor centers. This possibility is related to the circumstance that the fixed charges are not properly considered as trapped carriers, since the trapping processes are manifest through changes in fixed-charge concentrations rather than in these concentrations themselves. For example, centers of the acceptor type function as electron or hole traps according to whether the concentration of the charged centers increases or decreases with carrier injection. Consistent with the discussion in Section 2.1.2 of the steady-state trapping ratios, either type of center may be considered alternatively as an electron trap or a hole trap, under the convention that a change in fixed-charge concentration resulting from trapping may be negative as well as positive. To establish this result from (37), write the two equations that apply for, say, acceptor centers only. Then transform these so that the concentration $\mathfrak{N}_1 - \hat{n}$ of neutral centers becomes concentration of fixed positive charges $\hat{p}$, and the concentration $\hat{n}$ of charged centers becomes concentration $\mathfrak{N}_2 - \hat{p}$ of neutral centers; note that a given increase in the original $\hat{n}$ is equivalent to the corresponding decrease in the new $\hat{p}$. New equations then result that (with the replacement of C_{n1} and C_{p1} respectively by C_{n2} and C_{p2}) are the ones that follow directly from (37) for donor centers only.

2.2.1.1 *Thermal-Equilibrium Relationships.* The definitions

$$\begin{aligned} n_1 &\equiv n_0(\mathfrak{N}_1 - \hat{n}_0)/\hat{n}_0\,, & p_1 &\equiv p_0\hat{n}_0/(\mathfrak{N}_1 - \hat{n}_0), \\ n_2 &\equiv n_0\hat{p}_0/(\mathfrak{N}_2 - \hat{p}_0), & p_2 &\equiv p_0(\mathfrak{N}_2 - \hat{p}_0)/\hat{p}_0 \end{aligned} \tag{38}$$

are required by detailed balance. It is evident from these equations that

$$n_1 p_1 = n_2 p_2 = n_0 p_0 = n_i^2 \tag{39}$$

holds, where n_i is the thermal-equilibrium electron or hole concentration in intrinsic material. Note that (39) states, in effect, that the product — $(C_{n1}n_1)(C_{p1}p_1)$ or $(C_{n2}n_2)(C_{p2}p_2)$ — of the electron and hole emission coefficients equals n_i^2 times the product of the corresponding capture coefficients.[2,22,28]

If the concentrations in the right-hand members of (38) are known, then n_1, p_1, n_2 and p_2 are, of course, determined. Certain relationships hold between the concentrations. Since $n_0 - p_0 \equiv n_s$ equals $\hat{p}_0 - \hat{n}_0$ from (1), the neutrality condition, this condition and the last equation of (39) give

$$n_0 = \tfrac{1}{2}\{[(\hat{p}_0 - \hat{n}_0)^2 + 4n_i^2]^{\frac{1}{2}} + (\hat{p}_0 - \hat{n}_0)\}$$

and (40)

$$p_0 = \tfrac{1}{2}\{[(\hat{p}_0 - \hat{n}_0)^2 + 4n_i^2]^{\frac{1}{2}} - (\hat{p}_0 - \hat{n}_0)\}.$$

It is readily found from (38) that fractions of charged acceptor and donor centers are given respectively by

$$\hat{n}_0/\mathfrak{N}_1 = (1 + \alpha_{10}^{-1})^{-1}, \qquad \hat{p}_0/\mathfrak{N}_2 = (1 + \alpha_{20}^{-1})^{-1}, \tag{41}$$

with

$$\begin{aligned} \alpha_{10} &\equiv n_0/n_1 = p_1/p_0 = \tfrac{1}{2}n_1^{-1}[(n_s^2 + 4n_i^2)^{\frac{1}{2}} + n_s], \\ \alpha_{20} &\equiv p_0/p_2 = n_2/n_0 = \tfrac{1}{2}p_2^{-1}[(n_s^2 + 4n_i^2)^{\frac{1}{2}} - n_s]; \end{aligned} \tag{42}$$

the final expressions on the right follow by use of (40). For given semiconductor material at given temperature, n_i is known and n_s is determined by conductivity type and conductivity σ_0, and n_1 and n_2 (and hence p_1 and p_2 also) are accordingly determined by the fractions of charged centers. Expressing the thermal-equilibrium concentrations of mobile carriers and fixed charges in terms of each other (with other concentrations as parameters) thus involves roots of quadratic equations. The relationships given apply regardless of the number of kinds of centers present, since (40) contains no quantities pertaining to particular centers, and each equation of (41) and (42) contains quantities pertaining only to a single kind of center.

On the other hand, the fixed-charge and mobile-carrier concentrations for centers of two kinds are obtainable in general in terms of $\mathfrak{N}_1$, $\mathfrak{N}_2$, n_1, n_2 and n_i^2. It will suffice to indicate that the concentrations are roots of

biquadratics that follow readily from the equations

$$\begin{aligned} m_0 &= \hat{n}_0 + n_1\hat{n}_0/(\mathfrak{N}_1 - \hat{n}_0) = \hat{p}_0 + p_1(\mathfrak{N}_1 - \hat{n}_0)/\hat{n}_0 , \\ m_0 &= \hat{n}_0 + n_2(\mathfrak{N}_2 - \hat{p}_0)/\hat{p}_0 = \hat{p}_0 + p_2\hat{p}_0/(\mathfrak{N}_2 - \hat{p}_0) \end{aligned} \tag{43}$$

for $\hat{n}_0$ and $\hat{p}_0$ and

$$\begin{aligned} m_0 &= \alpha_{10}[n_1 + \mathfrak{N}_1/(1 + \alpha_{10})] = \alpha_{20}[p_2 + \mathfrak{N}_2/(1 + \alpha_{20})], \\ \alpha_{10}\alpha_{20} &= n_2/n_1 = p_1/p_2 \end{aligned} \tag{44}$$

for α_{10} and α_{20}, and hence n_0 and p_0. Equations (43) and (44) are obtained from (38) by eliminating, respectively, the concentrations of mobile carriers and fixed charges by use of the neutrality condition.† They are equivalent to combining (40) with (41) and (42), which are accordingly subject to a requirement of mutual consistency. For example, temperature determines n_i^2 for a given semiconductor; specifying conductivity also then determines n_0 and p_0 ; specifying further n_1 and n_2 determines $\hat{n}_0/\mathfrak{N}_1$ and $\hat{p}_0/\mathfrak{N}_2$ from (41) and (42), but only one of $\mathfrak{N}_1$ and $\mathfrak{N}_2$ can now be independently specified, since $\hat{p}_0 - \hat{n}_0$ must equal $n_0 - p_0$.

Through familiar considerations involving equilibrium Boltzmann statistics, the concentration n_1 or p_1 (and n_2 or p_2) has been shown to equal electron concentration in the conduction band or hole concentration in the valence band for the Fermi level coincident with the energy level of the centers.[31] The relationship

$$n_1 = n_i^2/p_1 = n_i e^{(\mathcal{E}_1 - \mathcal{E})/kT} = n_i e^{[e(\Phi - \Psi_1)/kT]} \tag{45}$$

for acceptor centers is here employed, and a similar one is used for donor centers. Here $\Psi_1 \equiv -e^{-1}\mathcal{E}_1$ and $\Phi \equiv -e^{-1}\mathcal{E}$ are the equivalent electrostatic potentials of the energy level $\mathcal{E}_1$ of the centers and the Fermi energy $\mathcal{E}$ for intrinsic material. This relationship is more phenomenological than those involving the energies of the conduction-band and valence-band edges and which give n_1 and p_1 in units of the effective densities of states in the bands. Note that the temperature dependence of the energy gap is involved through n_i, while the difference between the effective densities of states or the effective masses with nonspherical energy surfaces in momentum space is reflected simply in a difference between Φ and the midgap potential. If statistical weights associated with spin degeneracy are taken into account, then the definitions of (38) are of course retained, but (45) is modified. The right-hand members (for n_1) are multiplied

† It is easily seen that cubics result for centers of one kind only, or if complete ionization obtains for one of two kinds of centers.

by two; the exponentials for p_1 are multiplied by one-half. In the similar result for donor centers, the exponentials for n_2 and p_2 are multiplied by one-half and two, respectively. For given n_1 and n_2, these modifications [17,68,69] produce comparatively minor changes in $\mathcal{E}_1$ and $\mathcal{E}_2$ or Ψ_1 and Ψ_2.

2.2.1.2 *Equations in Concentration Increments; Trapping and Release Frequencies and Times.* For detailed analysis, it is advantageous to replace (37) by equations in the increments Δm, $\Delta\hat{n}$ and $\Delta\hat{p}$ in m, $\hat{n}$ and $\hat{p}$ over their thermal-equilibrium values. By subtracting from (37) the corresponding thermal-equilibrium equations, in which the time derivatives and the quantities in the various square brackets are zero, the result

$$\begin{aligned}
\Delta g - \Delta\mathfrak{R}_m &= \Delta g + \nu_{11}\Delta m + \nu_{12}\Delta\hat{n} + \nu_{13}\Delta\hat{p} - C_i\Delta n\Delta p \\
&\quad - C_{p1}\Delta p\Delta\hat{n} - C_{n2}\Delta n\Delta\hat{p}, \\
\partial\Delta\hat{n}/\partial t &= \Delta\mathfrak{R}_n - \Delta\mathfrak{R}_m = \nu_{21}\Delta m + \nu_{22}\Delta\hat{n} + \nu_{23}\Delta\hat{p} \\
&\quad - (C_{n1}\Delta n + C_{p1}\Delta p)\Delta\hat{n}, \\
\partial\Delta\hat{p}/\partial t &= \Delta\mathfrak{R}_p - \Delta\mathfrak{R}_m = \nu_{31}\Delta m + \nu_{32}\Delta\hat{n} + \nu_{33}\Delta\hat{p} \\
&\quad - (C_{n2}\Delta n + C_{p2}\Delta p)\Delta\hat{p}
\end{aligned} \tag{46}$$

follows, in which the decay constants of what will be referred to as the "ν_{ij} notation" are given by

$$\begin{aligned}
\nu_{11} &\equiv -C_i(n_0 + p_0) - C_{p1}\hat{n}_0 - C_{n2}\hat{p}_0, \\
\nu_{12} &\equiv C_i p_0 - C_{p1}(p_0 + p_1) + C_{n2}\hat{p}_0, \\
\nu_{13} &\equiv C_i n_0 - C_{n2}(n_0 + n_2) + C_{p1}\hat{n}_0, \\
\nu_{21} &\equiv C_{n1}(\mathfrak{N}_1 - \hat{n}_0) - C_{p1}\hat{n}_0, \\
\nu_{22} &\equiv -C_{n1}(\mathfrak{N}_1 - \hat{n}_0 + n_0 + n_1) - C_{p1}(p_0 + p_1), \\
\nu_{23} &\equiv C_{p1}\hat{n}_0, \\
\nu_{31} &\equiv C_{p2}(\mathfrak{N}_2 - \hat{p}_0) - C_{n2}\hat{p}_0, \\
\nu_{32} &\equiv C_{n2}\hat{p}_0, \\
\nu_{33} &\equiv -C_{n2}(n_0 + n_2) - C_{p2}(\mathfrak{N}_2 - \hat{p}_0 + p_0 + p_2).
\end{aligned} \tag{47}$$

Zero subscripts denote thermal equilibrium values. Note that $\Delta\mathfrak{R}_m$, $\Delta\mathfrak{R}_n$ and $\Delta\mathfrak{R}_p$ are respectively $\mathfrak{R}_m$, $\mathfrak{R}_n$ and $\mathfrak{R}_p$ minus $g_0 = C_i n_i^2$. In (46), in which Δm, $\Delta\hat{n}$ and $\Delta\hat{p}$ are to be considered as dependent variables, the

quadratic terms have been written compactly with Δn and Δp, which may be replaced by $\Delta m - \Delta\hat{n}$ and $\Delta m - \Delta\hat{p}$; and n_0 and p_0 in (47) may be replaced by $m_0 - \hat{n}_0$ and $m_0 - \hat{p}_0$.

It is desirable to supplement the ν_{ij} notation with another notation, which, although it often results in less compact expressions, facilitates physical interpretations. If volume generation and direct recombination are neglected for the present, the respective contributions to $\partial\Delta n/\partial t$ and $\partial\Delta p/\partial t$ other than the terms involving transport processes as such may be written as

$$\begin{aligned} -\Delta\mathfrak{R}_n = -\Delta\mathfrak{R}_m - \partial\Delta\hat{n}/\partial t = &- \nu_{tn1}\Delta n + \nu_{gn1}\Delta\hat{n} \\ &- \nu_{tn2}\Delta n + \nu_{gn2}\Delta(\mathfrak{N}_2 - \hat{p}), \\ -\Delta\mathfrak{R}_p = -\Delta\mathfrak{R}_m - \partial\Delta\hat{p}/\partial t = &- \nu_{tp1}\Delta p + \nu_{gp1}\Delta(\mathfrak{N}_1 - \hat{n}) \\ &- \nu_{tp2}\Delta p + \nu_{gp2}\Delta\hat{p}. \end{aligned} \tag{48}$$

The top and bottom rows of the forms on the right give the respective contributions of the acceptor and donor centers. The decay constants may be identified as certain capture and release frequencies by comparison with the equations

$$\begin{aligned} -\Delta\mathfrak{R}_n = &- C_{n1}[(\mathfrak{N}_1 - \hat{n}_0)\Delta n - n_0\Delta\hat{n} - \Delta n\Delta\hat{n}] + C_{n1}n_1\Delta\hat{n} \\ &- C_{n2}[\hat{p}_0\Delta n + n_0\Delta\hat{p} + \Delta n\Delta\hat{p}] + C_{n2}n_2\Delta(\mathfrak{N}_2 - \hat{p}), \\ -\Delta\mathfrak{R}_p = &- C_{p1}[\hat{n}_0\Delta p + p_0\Delta\hat{n} + \Delta p\Delta\hat{n}] + C_{p1}p_1\Delta(\mathfrak{N}_1 - \hat{n}) \\ &- C_{p2}[(\mathfrak{N}_2 - \hat{p}_0)\Delta p - p_0\Delta\hat{p} - \Delta p\Delta\hat{p}] + C_{p2}p_2\Delta\hat{p}, \end{aligned} \tag{49}$$

which follow from (37). In (49), the magnitudes of the contributions involving brackets are capture rates, while the other terms on the right are release rates.

Expression of the capture rates in terms of capture frequencies would require writing them with Δn or Δp as a factor, and would thus necessitate solution of the particular problem. These physical capture frequencies would depend in general on coordinates and time. The contributions to the capture rates that contain $\Delta\hat{n}$ and $\Delta\hat{p}$ as factors are associated, however, with trap saturation: These contributions, for carriers of given charge, represent the decreases and increases in capture rate with the filling of centers that assume, respectively, the same and the opposite charges. They may, in a phenomenological sense, be deleted from the capture rates and assigned to the release rates. The "effective" capture and release rates that result from this procedure are clearly rates in

terms of which the capture and release frequencies of (48) and related times may be defined as follows:

Electron capture by neutral acceptors:

$$\nu_{tn1} \equiv \tau_{tn1}^{-1} \equiv C_{n1}(\mathfrak{N}_1 - \hat{n}_0) = C_{n1}\mathfrak{N}_1/(1 + \alpha_{10}),$$

Electron release from charged acceptors:

$$\nu_{gn1} \equiv \tau_{gn1}^{-1} \equiv C_{n1}(n + n_1),$$

Hole capture by charged acceptors:

$$\nu_{tp1} \equiv \tau_{tp1}^{-1} \equiv C_{p1}\hat{n}_0 = C_{p1}\mathfrak{N}_1\alpha_{10}/(1 + \alpha_{10}),$$

Hole release from neutral acceptors:

$$\nu_{gp1} \equiv \tau_{gp1}^{-1} \equiv C_{p1}(p + p_1), \tag{50}$$

Electron capture by charged donors:

$$\nu_{tn2} \equiv \tau_{tn2}^{-1} \equiv C_{n2}\hat{p}_0 = C_{n2}\mathfrak{N}_2\alpha_{20}/(1 + \alpha_{20}),$$

Electron release from neutral donors:

$$\nu_{gn2} \equiv \tau_{gn2}^{-1} \equiv C_{n2}(n + n_2),$$

Hole capture by neutral donors:

$$\nu_{tp2} \equiv \tau_{tp2}^{-1} \equiv C_{p2}(\mathfrak{N}_2 - \hat{p}_0) = C_{p2}\mathfrak{N}_2/(1 + \alpha_{20}),$$

Hole release from charged donors:

$$\nu_{gp2} \equiv \tau_{gp2}^{-1} \equiv C_{p2}(p + p_2).$$

The second forms given for capture frequencies follow by use of (41). Note, for example, that ν_{tn1} is an average frequency per electron of electron capture by a neutral acceptor center and hence the reciprocal of the corresponding electron capture or trapping time, τ_{tn1} ; and that ν_{gn1} is an average frequency per charged center of electron release from a charged acceptor center and hence the reciprocal of the corresponding electron release time, τ_{gn1} . The saturation terms that originate from the true capture rates appear as the contributions from n and p in the "effective" release frequencies, while the "effective" capture frequencies do not depend on the injection level.

It is readily seen that, if direct recombination is neglected, then an alternative procedure† for including the quadratic terms in (46) is to

† Another alternative procedure is to replace $\hat{n}_0$ by $\hat{n}$ and $\hat{p}_0$ by $\hat{p}$ or, more generally, by increasing $\hat{n}_0$ and $\hat{p}_0$ by a fraction γ of $\Delta\hat{n}$ and of $\Delta\hat{p}$ and n_0 and p_0 by the fraction $1 - \gamma$ of Δn and of Δp. The definitions of (50), which correspond to $\gamma = 0$, are then modified, and depend on γ, the fraction of the quadratic terms assigned to capture.

generalize the ν_{ij} by replacing n_0 and p_0 in (47) by n and p. The formal equivalence of (46) and (48) that results then permits expressing these generalized ν_{ij} in terms of the effective capture and release frequencies of (50), as follows:

$$\begin{aligned}
\nu_{11} &= -\nu_{tp1} - \nu_{tn2}\,, \\
\nu_{12} &= -\nu_{gp1} + \nu_{tn2}\,, \\
\nu_{13} &= \nu_{tp1} - \nu_{gn2}\,, \\
\nu_{21} &= \nu_{tn1} - \nu_{tp1}\,, \\
\nu_{22} &= -\nu_{tn1} - \nu_{gn1} - \nu_{gp1}\,, \\
\nu_{23} &= \nu_{tp1}\,, \\
\nu_{31} &= -\nu_{tn2} + \nu_{tp2} \\
\nu_{32} &= \nu_{tn2}\,, \\
\nu_{33} &= -\nu_{gn2} - \nu_{tp2} - \nu_{gp2}\,.
\end{aligned} \tag{51}$$

Note that

$$\nu_{11} + \nu_{23} + \nu_{32} = 0 \tag{52}$$

holds for this case of no direct recombination.

The four effective trapping and release times or frequencies for each type of center satisfy a fundamental restriction, namely:

$$\frac{\tau_{gnj}\tau_{tpj}}{\tau_{tnj}\tau_{gpj}} = \frac{\nu_{tnj}\nu_{gpj}}{\nu_{gnj}\nu_{tpj}} = \frac{p_0}{n_0}\frac{1 + \Delta p/(p_0 + p_j)}{1 + \Delta n/(n_0 + n_j)}, \qquad j = 1, 2. \tag{53}$$

Thus, only three are independent. As will appear, this restriction is widely useful for calculations and physical interpretations. It is essentially a consequence of detailed balance: For thermal equilibrium, it follows readily from relationships tantamount to this principle, such as (41) and (42) or the definitions of (38). The factor on the right that depends on Δn and Δp results simply from the concentration dependence of the effective release frequencies.

A property easily established from (42) and (50) is the following: For centers of given capture coefficients and energy level, if the electron and hole capture frequencies are equal for a given conductivity, then the equilibrium release frequencies are equal for material of the opposite conductivity type and the same value of $|\, n_0 - p_0 \,|$, that is, for material such that the values of n_0 and p_0 are, in effect, interchanged.

2.2.1.3 *Trapping and Recombination Ranges; Shallow and Deep Traps.* Three linear small-signal ranges, characterized respectively primarily by minority-carrier trapping, recombination and majority-carrier trapping, may be defined for each type of center by use of (53). The "minority-carrier trapping range" is defined by the condition that the equilibrium minority-carrier to majority-carrier release frequency ratio exceeds unity. In p-type material, this ratio, ν_{gnj}/ν_{gpj}, is $C_{nj}n_j/C_{pj}p_0 = C_{nj}n_0/C_{pj}p_j$, from (39), (41), (42) and (50); and, from (53), ν_{tnj}/ν_{tpj} is larger by the factor p_0/n_0. The "majority-carrier trapping range" is defined by the condition that the majority- to minority-carrier capture frequency ratio exceeds unity, for which the equilibrium majority- to minority-carrier release frequency ratio is larger by the factor p_0/n_0 for p-type material, or by n_0/p_0 for n-type. The "recombination range" is defined as that not included in either trapping range. Thus, the recombination range is given by $n_0/n_j = p_j/p_0 \leqq C_{nj}/C_{pj} \leqq p_j/n_0 = p_0/n_j$ for p-type material, the electron-trapping range by $C_{nj}/C_{pj} > p_j/n_0 = p_0/n_j$, and the hole-trapping range by $C_{nj}/C_{pj} < p_j/p_0 = n_0/n_j$. A "minority-carrier capture range", which includes the trapping and recombination ranges, may be defined by $\nu_{tnj}/\nu_{tpj} > 1$. Similar results, obtainable by interchanging n and p, hold for n-type material. Ranges of minority-carrier-dominated and majority-carrier-dominated transitions[22,28] are those parts of the trapping ranges here considered for which strong inequalities hold. Equal capture frequencies, which occur at the boundary between the recombination and majority-carrier trapping ranges, result in what will be termed "critical recombination", with which, as will be seen, $\Delta\hat{n}$ or $\Delta\hat{p}$ is zero.

The three ranges may be specified in terms of the equality densities. These are the equilibrium carrier concentrations for the Fermi level coincident with the equality level. They are defined in the present context by

$$n_j^* \equiv C_{pj}p_j/C_{nj} = p_0\nu_{tpj}/\nu_{tnj} = n_0\nu_{gpj}/\nu_{gnj}$$

and (54)

$$p_j^* \equiv C_{nj}n_j/C_{pj} = n_0\nu_{tnj}/\nu_{tpj} = p_0\nu_{gnj}/\nu_{gpj},$$

in which the release frequencies are equilibrium values. Thus, the recombination range is given by $n_0 \leqq p_j^* \leqq p_0$ or $p_0 \geqq n_j^* \geqq n_0$ for p-type material, the electron-trapping range by $n_j^* < n_0$ or $p_j^* > p_0$ and the hole-trapping range by $n_j^* > p_0$ or $p_j^* < n_0$, and similarly for n-type material. The ranges may evidently also be specified in terms of

the equality level, the Fermi level $\mathcal{E}$ for intrinsic material, the actual Fermi level $\mathcal{E}_F$ and the "reflected Fermi level," $\mathcal{E}_F' \equiv 2\mathcal{E} - \mathcal{E}_F$, the reflection of $\mathcal{E}_F$ about $\mathcal{E}$: For the recombination range, the equality level is between $\mathcal{E}_F$ and $\mathcal{E}_F'$; for the minority-carrier trapping range, it is between $\mathcal{E}_F$ and the edge of the majority-carrier band; and for the majority-carrier trapping range, it is between $\mathcal{E}_F'$ and the edge of the minority-carrier band. Note that, if the capture coefficients are equal, then $n_j^* = p_j$ (or $p_j^* = n_j$) holds and the respective trapping ranges are given by conditions on the trapping level $\mathcal{E}_j$ obtained by interchanging those on the equality level.

The volume rates of electron and of hole transitions at equilibrium are respectively $C_{n1}n_0(\mathfrak{N}_1 - \hat{n}_0) = C_{n1}n_1\hat{n}_0 = n_0\nu_{tn1}$ and $C_{p1}p_0\hat{n}_0 = C_{p1}p_1(\mathfrak{N}_1 - \hat{n}_0) = p_0\nu_{tp1}$ for acceptor-type centers. From (53), these rates are proportional to† ν_{gn1} and ν_{gp1}. Hence each definition given for a trapping range insures that the transition rate at equilibrium for the particular carriers is the larger, and also that the transition rate ν_{tn1} or ν_{tp1} per mobile carrier is the larger too. The asymmetrical relationship between the definitions for minority- and majority-carrier trapping reflects the circumstance that a transition rate will be the larger if either the cross section or the concentration of the particular carriers is sufficiently large. The recombination range is that for which a larger transition rate per mobile minority carrier is associated with a total transition rate for majority carriers which is the larger.

For shallow minority-carrier traps, since relatively few are occupied by minority carriers at equilibrium so that they can capture majority carriers, the condition for the minority-carrier trapping range may be met even though the capture coefficients are comparable in magnitude. For deep traps, since relatively few can capture minority carriers, the minority-carrier trapping generally requires a minority-carrier capture coefficient considerably the larger. Suitable condition for "shallow" and "deep" traps are, in view of the condition on C_{nj}/C_{pj} for the electron-trapping range, respectively $p_j \ll n_0$ (or $n_j \gg p_0$) and $n_j \ll p_0$ (or $p_j \gg n_0$) in p-type material. That is, "shallow" and "deep" traps for minority carriers are appreciably removed from the reflected Fermi level $\mathcal{E}_F'$, towards the edges of the minority- or majority-carrier band. Similarly, for majority-carrier trapping, "shallow" and "deep" traps are appreciably removed from the Fermi level $\mathcal{E}_F$, towards the edges of the majority- or minority-carrier band, respectively.

† They equal ν_{gn1} and ν_{gp1} times the capture concentration, as shown by (63) in Section 3.1.1.

2.2.2 *Centers with Two Energy Levels*

The formalism for centers of two types is readily modified to yield equations for one type of center with two energy levels. With the assumption that the centers can each assume single negative or positive charge or be neutral, $\hat{n}$ and $\hat{p}$ denote concentrations of centers in the respective charged states. It is thus clear that the fundamental mass-action equations for this case are formally the same as (37), with the modification that both $\mathfrak{N}_1 - \hat{n}$ and $\mathfrak{N}_2 - \hat{p}$ are replaced by $\mathfrak{N} - \hat{n} - \hat{p}$, where $\mathfrak{N}$ is the total concentration of the centers.

For thermal equilibrium, definitions of n_1, p_1, n_2 and p_2 apply that are equations of (38) with both $\mathfrak{N}_1 - \hat{n}_0$ and $\mathfrak{N}_2 - \hat{p}_0$ replaced by $\mathfrak{N} - \hat{n}_0 - \hat{p}_0$. It follows that the restriction

$$n_0^2/n_1n_2 = p_1p_2/p_0^2 = \hat{n}_0/\hat{p}_0 \tag{55}$$

holds for this two-level case. As is easily verified, (39) and (40) still apply, while the fractions of charged centers are

$$\begin{aligned}
\hat{n}_0/\mathfrak{N} &= (1 + n_1/n_0 + n_1n_2/n_0^2)^{-1} \\
&= (1 + p_0/p_1 + p_0^2/p_1p_2)^{-1} \\
&= \alpha_{10}/(1 + \alpha_{10} + \alpha_{20}), \\
\hat{p}_0/\mathfrak{N} &= (1 + p_2/p_0 + p_1p_2/p_0^2)^{-1} \\
&= (1 + n_0/n_2 + n_0^2/n_1n_2)^{-1} \\
&= \alpha_{20}/(1 + \alpha_{10} + \alpha_{20}),
\end{aligned} \tag{56}$$

with α_{10} and α_{20} given by (42). The modifications of (43) and (44) for the biquadratics are the replacement of $\mathfrak{N}_1 - \hat{n}_0$ and $\mathfrak{N}_2 - \hat{p}_0$ by $\mathfrak{N} - \hat{n}_0 - \hat{p}_0$ and of $\mathfrak{N}_1/(1 + \alpha_{10})$ and $\mathfrak{N}_2(1 + \alpha_{20})$ by $\mathfrak{N}/(1 + \alpha_{10} + \alpha_{20})$. Note that (55) is not an independent equation, in that it is implicit in the modified (43).† Relationships formally identical with (45) give n_1 and n_2 in terms of the two energy levels.

The fundamental two-level mass-action equations for no direct recombination yield the equations in concentration increments

$$\begin{aligned}
\Delta g - \Delta \mathfrak{R}_m &= \Delta g - (\nu_{tp1} + \nu_{tn2})\Delta m + (\nu_{tn2} - \nu_{gp1} - C_{n2}n_2)\Delta\hat{n} \\
&\quad + (\nu_{tp1} - \nu_{gn2} - C_{p1}p_1)\Delta\hat{p}, \\
\partial\Delta\hat{n}/\partial t &= (\nu_{tn1} - \nu_{tp1})\Delta m - (\nu_{tn1} + \nu_{gn1} + \nu_{gp1})\Delta\hat{n} \\
&\quad + (\nu_{tp1} - C_{n1}n - C_{p1}p_1)\Delta\hat{p}, \\
\partial\Delta\hat{p}/\partial t &= (\nu_{tp2} - \nu_{tn2})\Delta m + (\nu_{tn2} - C_{p2}p - C_{n2}n_2)\Delta\hat{n} \\
&\quad - (\nu_{tp2} + \nu_{gp2} + \nu_{gn2})\Delta\hat{p},
\end{aligned} \tag{57}$$

† The corresponding restriction for two types of centers has $(\hat{n}_0/\hat{p}_0)\ (\mathfrak{N}_2 - \hat{p}_0)/(\mathfrak{N}_1 - \hat{n}_0)$ as right-hand member.

from which appropriate ν_{ij} can immediately be identified. Effective capture and release frequencies are here employed whose definitions are provided by (50) if $\mathfrak{N}_1 - \hat{n}_0$ and $\mathfrak{N}_2 - \hat{p}_0$ are replaced by $\mathfrak{N} - \hat{n}_0 - \hat{p}_0$, and $\mathfrak{N}_1/(1 + \alpha_{10})$ and $\mathfrak{N}_2/(1 + \alpha_{20})$ replaced by $\mathfrak{N}_2/(1 + \alpha_{10} + \alpha_{20})$. Aside from these modified definitions, the equations of (57) are formally identical with equations for two kinds of centers except for the additional terms in which the capture coefficients appear. These are "constraint" terms. The ones in $\partial\Delta\hat{n}/\partial t$ represent the decrease in this rate that results from the decrease in the concentration of neutral centers associated with an increase in $\hat{p}$; neutral centers capturing electrons and emitting holes are the two processes that increase $\hat{n}$. The rate decrease $C_{n1}n\Delta\hat{p}$ is that associated with the electron capture, while $C_{p1}p_1\Delta\hat{p}$ is that associated with the hole emission. The condition that the rate decreases for these two processes be the same is clearly $n_1^* = n$. Similarly, the constraint terms in $\partial\Delta\hat{p}/\partial t$ represent the respective decreases $C_{p2}p\Delta\hat{n}$ and $C_{n2}n_2\Delta\hat{n}$ in the neutral-center hole capture and electron emission rates associated with an increase in $\hat{n}$; these decreases are equal if $p_2^* = p$ holds. The third forms of (54) show that a pair of equal constraint terms implies an equilibrium hole-to-electron release-frequency or transition-rate ratio for the acceptor or donor levels equal, respectively, to n/n_0 or p_0/p, which are substantially unity near thermal equilibrium.

For this two-level case, the four effective trapping and release times or frequencies associated with each energy level satisfy the fundamental restriction that is formally identical with (53). It is also easily verified that the various conditions given for the recombination and trapping ranges and for shallow and deep traps apply without formal modification.

By suitable notational generalization of the fundamental mass-action equations, the results of this section can be shown to apply to two-level centers in general, whose states (differing successively by one electronic charge) may include ones that are multiply charged, either positively or negatively. Through use of the phenomenological capture coefficients, statistical weights associated with multiply charged states do not enter explicitly. For example, the concentration $\mathfrak{N} - \hat{n} - \hat{p}$ of neutral centers may be replaced by concentration $\tilde{p}$ of centers with single positive charge, and $\hat{p}$ used to denote concentration of centers with double positive charge. Then $\hat{n}$ is replaced by the new concentration $\mathfrak{N} - \tilde{p} - \hat{p}$ of neutral centers.† Thus, with obvious modifications in the physical descrip-

† Note that these transformations applied to (1) give $\Delta m = \Delta n - \Delta\tilde{p} - \Delta\hat{p} = \Delta p + \Delta\hat{p}$. While the correct neutrality condition holds, Δm is no longer the increment in concentration of total negative or positive charges.

tion of capture and release frequencies and other quantities, the theory is essentially unchanged.

2.2.3 *Volume Generation with Excitations Involving Trapping Levels*

The excitations associated with the absorption of radiation of wavelengths beyond the limit of intrinsic absorption may be taken into account phenomenologically through suitable generation terms in the differential equations. To the volume rate g of interband excitations in the differential equation for n is added $g_{c1} + g_{c2}$, where g_{c1} is the volume rate of hole excitations from the conduction band to centers of type 1 — that is, electron excitations from these centers to the conduction band — and g_{c2} is the similar quantity for centers of type 2. Similarly, in the differential equation for p, to g is added $g_{v1} + g_{v2}$, each term of which is the volume rate of electron excitations from the valence band to the centers or hole excitations from the centers to the valence band. To g in the differential equation for m is added $g_{v1} + g_{c2}$, and not g_{c1} or g_{v2}, since g_{c1} increases n as it decreases $\hat{n}$, while g_{v2} increases p as it decreases $\hat{p}$. The generation terms Δg_m, Δg_n and Δg_p in the differential equations for Δm, Δn and Δp are thus†

$$\begin{aligned} \Delta g_m &\equiv \Delta g + \Delta g_{v1} + \Delta g_{c2}, \\ \Delta g_n &\equiv \Delta g + \Delta g_{c1} + \Delta g_{c2}, \\ \Delta g_p &\equiv \Delta g + \Delta g_{v1} + \Delta g_{v2}, \end{aligned} \tag{58}$$

and the generation terms that the equations for $\Delta\hat{n}$ and $\Delta\hat{p}$ now contain are respectively $\Delta g_m - \Delta g_n = \Delta g_{v1} - \Delta g_{c1}$ and $\Delta g_m - \Delta g_p = \Delta g_{c2} - \Delta g_{v2}$.

The additional generation terms clearly represent the same processes as do the emission terms of (37). The distinction implicit in the notation is valid, however, consistent with zero values of these additional generation terms at thermal equilibrium. Each generation rate of (37) is determined at equilibrium by the phonons and radiation associated with the equal corresponding capture rate. Since detailed balance applies also to the radiative part separately, there is no net radiation at equilibrium from any given process of capture and the corresponding generation.

III. DETAILED THEORY AND APPLICATIONS

3.1 *Diffusion Length and Steady-State Lifetime Functions*

3.1.1 *Linear Theory*

The equations of (46) for two types of centers, when written for the steady state and linearized by neglect of the quadratic terms, give con-

† The excitations involving trapping levels only, which may occur for large concentrations of centers (presumably with concomitant impurity-band conduction), are here neglected.

centration increments that are proportional, and solving for $\Delta\hat{n}/\Delta m$ and $\Delta\hat{p}/\Delta m$ provides the thermal-equilibrium trapping ratios. These and the corresponding lifetime τ_m are thus given by

$$\begin{aligned} r_n &= \frac{\nu_{23}\nu_{31} - \nu_{21}\nu_{33}}{\nu_{22}\nu_{33} - \nu_{23}\nu_{32}}, \\ r_p &= \frac{\nu_{21}\nu_{32} - \nu_{22}\nu_{31}}{\nu_{22}\nu_{33} - \nu_{23}\nu_{32}}, \\ \tau_m &= -(\nu_{11} + \nu_{12}r_n + \nu_{13}r_p)^{-1} \\ &= \frac{-(\nu_{22}\nu_{33} - \nu_{23}\nu_{32})}{\nu_{11}(\nu_{22}\nu_{33} - \nu_{23}\nu_{32}) + \nu_{12}(\nu_{23}\nu_{31} - \nu_{21}\nu_{33}) + \nu_{13}(\nu_{21}\nu_{32} - \nu_{22}\nu_{31})}, \end{aligned} \tag{59}$$

in terms of which, with the thermal-equilibrium diffusivity D_0' from (31), the diffusion length can be expressed and the diffusion-length lifetime evaluated. These results apply also for two-level centers if equilibriums ν_{ij} are defined in accordance with (57).

The case of single-level centers of one type lends itself to more detailed analysis. Results from the linearized equations for, say, acceptor centers only, for which r_p is zero, are:

$$\begin{aligned} r_n &= -\nu_{21}/\nu_{22} = \frac{\nu_{tn1} - \nu_{tp1}}{\nu_{tn1} + \nu_{gn1} + \nu_{gp1}} \\ &= \frac{\tau_{p0}(\mathfrak{N}_1 - \hat{n}_0) - \tau_{n0}\hat{n}_0}{\tau_{p0}(\mathfrak{N}_1 - \hat{n}_0 + n_0 + n_1) + \tau_{n0}(p_0 + p_1)} \\ &= \frac{\mathfrak{N}_1^*(\tau_{tp1} - \tau_{tn1})}{(\mathfrak{N}_1^* + n_0)\tau_{tp1} + p_0\tau_{tn1}} \\ &= 1 - \tau_n/\tau_p, \\ \tau_n &= (1 - r_n)\tau_m = -(\nu_{21} + \nu_{22})/\Delta_1 \\ &= (\nu_{tp1} + \nu_{gp1} + \nu_{gn1})/\Delta_1 \\ &= \frac{\tau_{n0}(\hat{n}_0 + p_0 + p_1) + \tau_{p0}(n_0 + n_1)}{\mathfrak{N}_1^* + n_0 + p_0} \\ &= \frac{(\mathfrak{N}_1^* + p_0)\tau_{tn1} + n_0\tau_{tp1}}{\mathfrak{N}_1^* + n_0 + p_0}, \\ \tau_p &= \tau_m = -\nu_{22}/\Delta_1 \\ &= (\nu_{tn1} + \nu_{gn1} + \nu_{gp1})/\Delta_1 \\ &= \frac{\tau_{p0}(\mathfrak{N}_1 - \hat{n}_0 + n_0 + n_1) + \tau_{n0}(p_0 + p_1)}{\mathfrak{N}_1^* + n_0 + p_0} \\ &= \frac{(\mathfrak{N}_1^* + n_0)\tau_{tp1} + p_0\tau_{tn1}}{\mathfrak{N}_1^* + n_0 + p_0}. \end{aligned} \tag{60}$$

The ν_{ij} and release frequencies are equilibrium values; τ_{n0} and τ_{p0}, given by

$$\tau_{n0} \equiv (C_{n1}\mathfrak{N}_1)^{-1} = (1 - \hat{n}_0/\mathfrak{N}_1)\tau_{tn1} = (1 + p_1/p_0)^{-1}\tau_{tn1}$$

and (61)

$$\tau_{p0} \equiv (C_{p1}\mathfrak{N}_1)^{-1} = (\hat{n}_0/\mathfrak{N}_1)\tau_{tp1} = (1 + n_1/n_0)^{-1}\tau_{tp1},$$

are the respective limiting lifetimes† in strongly extrinsic p- and n-type materials (in which they are also τ_{tn1} and τ_{tp1}); Δ_1, given by

$$\begin{aligned}\Delta_1 \equiv \nu_{11}\nu_{22} - \nu_{12}\nu_{21} &= \nu_{tn1}\nu_{gp1} + \nu_{tn1}\nu_{tp1} + \nu_{tp1}\nu_{gn1} \\ &= C_{n1}C_{p1}\mathfrak{N}_1(\mathfrak{N}_1^* + n_0 + p_0),\end{aligned} \tag{62}$$

is always positive if neither C_{n1} nor C_{p1} is zero; and $\mathfrak{N}_1^*$, which will be referred to as the "capture concentration," is given variously by

$$\begin{aligned}\mathfrak{N}_1^* &\equiv \nu_{tn1}\nu_{tp1}/C_{n1}C_{p1}\mathfrak{N}_1 = n_i^2\mathfrak{N}_1/(n_0 + n_1)(p_0 + p_1) \\ &= n_0\nu_{tn1}/\nu_{gn1} = p_0\nu_{tp1}/\nu_{gp1} = n_i(\tau_{gn1}\tau_{gp1}/\tau_{tn1}\tau_{tp1})^{\frac{1}{2}} \\ &= \alpha_{10}(\mathfrak{N}_1 - \hat{n}_0)/(1 + \alpha_{10}) = \hat{n}_0/(1 + \alpha_{10}) \\ &= \mathfrak{N}_1(\hat{n}_0/\mathfrak{N}_1)(1 - \hat{n}_0/\mathfrak{N}_1) \\ &= \mathfrak{N}_1\alpha_{10}/(1 + \alpha_{10})^2.\end{aligned} \tag{63}$$

The different forms for these results are obtained by use of (51) for the ν_{ij}, definitions of (50) for the capture and release frequencies, and equilibrium relationships of (38), (41), (42) and (53). The middle term of the second form for Δ_1 is the one that gives rise to $\mathfrak{N}_1^*$, and it follows from (53) that the first term is large or small compared with the third according to whether p_0 is large or small compared with n_0. Capture concentration $\mathfrak{N}_1^*$ large compared with $n_0 + p_0$ is, as will be shown in Section 3.2.1, the condition that capture frequencies predominate over release frequencies. The volume rates of electron and hole transitions at equilibrium (see Section 2.2.1.2), $n_0\nu_{tn1}$ and $p_0\nu_{tp1}$, are equal to $\mathfrak{N}_1^*$ times the corresponding release frequencies. The equations of (63) show that $\hat{n}_0$ and $\mathfrak{N}_1 - \hat{n}_0$ may be written in terms of $\mathfrak{N}_1^*$ and α_{10}. The concentration $\mathfrak{N}_1^*$ is small if the centers are nearly all ionized or un-ionized; the last form shows that its largest value is $\frac{1}{4}\mathfrak{N}_1$, which it assumes for $\alpha_{10} = 1$ or $\hat{n}_0/\mathfrak{N}_1 = \frac{1}{2}$, that is, for the Fermi level coincident with the energy level of the centers. Entirely similar results, for which obvious notational changes are required in some forms, hold for donor centers only.

† Conditions for these lifetimes are $p_0 \gg \mathfrak{N}_1^* + p_1 + p_1^*, n_0 \gg \mathfrak{N}_1^* + n_1 + n_1^*$.

The diffusion length L_0 and lifetime τ_0 corresponding to the electron and hole lifetimes of (60), which are the Shockley-Read lifetimes, may be evaluated from (31) or (35) and (60). These equations give

$$\begin{aligned} L_0^2 = D_0'\tau_m = D_0\tau_0 &= D_0[1 - r_n p_0/(n_0 + p_0)]\tau_p \\ &= kT\mu_n\mu_p\sigma_0^{-1}(n_0\tau_{tp1} + p_0\tau_{tn1}) \\ &= \sigma_0^{-1}(\sigma_{p0}D_n\tau_{tn1} + \sigma_{n0}D_p\tau_{tp1}), \end{aligned} \tag{64}$$

where σ_{n0} and σ_{p0} are $e\mu_n n_0$ and $e\mu_p p_0$. Other forms may be written by expressing τ_{tn1} and τ_{tp1} in terms of τ_{n0} and τ_{p0} by use of (61). The diffusion-length lifetime for this case,

$$\begin{aligned} \tau_0 &= (\nu_{gn1} + \nu_{gp1})/(\nu_{tn1}\nu_{gp1} + \nu_{tp1}\nu_{gn1}) \\ &= [\tau_{p0}(n_0 + n_1) + \tau_{n0}(p_0 + p_1)]/(n_0 + p_0) \\ &= (n_0\tau_{tp1} + p_0\tau_{tn1})/(n_0 + p_0), \end{aligned} \tag{65}$$

is formally similar to the familiar common lifetime[31,32] for both electrons and holes for the limiting case of $\mathfrak{N}_1$ small, as inspection of (60) serves to verify.† Thus, L_0 and τ_0 are, for given τ_{n0} and τ_{p0} or τ_{tn1} and τ_{tp1}, independent of $\mathfrak{N}_1$. For given energy level and capture coefficients, τ_0 is proportional to $\mathfrak{N}_1^{-1}$. The true L_0 and τ_0 apply, of course, in the linear part of the small-signal range, in which no appreciable trap saturation occurs. With small-signal trap saturation, diffusion length and lifetime that are usually considerably larger apply in the saturation range. These are evaluated in Section 3.1.2.

It can be shown that the electron and hole lifetimes of (60) are substantially equal to τ_0 if

$$|\nu_{tn1} - \nu_{tp1}| = |\nu_{gn1}/n_0 - \nu_{gp1}/p_0|\,\mathfrak{N}_1^* \ll (1 + \epsilon)(\nu_{gn1} + \nu_{gp1}) \tag{66}$$

holds, in which ϵ is the smaller of $(\mathfrak{N}_1^* + n_0)/p_0$ and $(\mathfrak{N}_1^* + p_0)/n_0$, as given by the respective conditions $|\tau_p - \tau_0|/\tau_0 \ll 1$ and $|\tau_n - \tau_0|/\tau_0 \ll 1$. For extrinsic material, ϵ may usually be neglected; including it provides an appreciably weaker condition only if $\mathfrak{N}_1^*$ is larger than the equilibrium majority-carrier concentration. The condition of (66) may be severe: It is essentially $\mathfrak{N}_1^*$ small compared with the equilibrium minority-carrier concentration for the minority- to majority-carrier release frequency ratio of order unity or larger in extrinsic material, that is, for the minority-carrier trapping range defined in Section 2.2.1.2.

† This formal similarity holds for any number M of kinds of centers, τ_0 being given by $[\sum_{j=1}^{M}(n_0\tau_{tpj} + p_0\tau_{tnj})^{-1}]^{-1}/(n_0 + p_0)$, as may be shown from the first form for $\mathfrak{R}_m$ of (71) and the observation that, if two or more different kinds of centers are present, then $\mathfrak{R}_m$ is the sum of similar terms for each kind.

General conditions for the validity of this linear analysis may be formulated as conditions for the neglect of the quadratic terms. For this purpose, assume uniform concentrations and volume-generation rate. Then (46) and (51) yield

$$d\Delta n/dt = \Delta g - \nu_{tn1}\Delta n + \nu_{gn1}\Delta\hat{n} + C_{n1}\Delta n\Delta\hat{n} = 0$$

and (67)

$$d\Delta p/dt = \Delta g - \nu_{tp1}\Delta p - \nu_{gp1}\Delta\hat{n} - C_{p1}\Delta p\Delta\hat{n} = 0$$

for acceptor centers only in the steady state and no direct recombination. The conditions may be derived in a self-consistent manner by first obtaining, with the neutrality condition, the concentrations from the linearized forms of (67). These concentrations, namely

$$\begin{aligned} \Delta n &= (\nu_{tp1} + \nu_{gp1} + \nu_{gn1})\Delta g/\Delta_1 , \\ \Delta\hat{n} &= (\nu_{tn1} - \nu_{tp1})\Delta g/\Delta_1 , \\ \Delta p &= (\nu_{tn1} + \nu_{gn1} + \nu_{gp1})\Delta g/\Delta_1 , \end{aligned} \tag{68}$$

are then substituted in (67), so that conditions for negligible quadratic terms may be obtained as restrictions on (positive) Δg and, by use of (68), as corresponding restrictions on the concentrations. It will suffice to give the former restrictions, which for the neglect of $C_{n1}\Delta n\Delta\hat{n}$ and $C_{p1}\Delta p\Delta\hat{n}$ are, respectively,

$$\Delta g/\Delta_1 \ll \nu_{tn1}/C_{n1}(\nu_{tn1} - \nu_{tp1}) = (\mathfrak{N}_1 - \hat{n}_0)/(\nu_{tn1} - \nu_{tp1})$$

and (69)

$$\Delta g/\Delta_1 \ll \Delta_1/C_{p1}(\nu_{tn1} - \nu_{tp1})(\nu_{tn1} + \nu_{gn1} + \nu_{gp1})$$

for $\nu_{tn1} > \nu_{tp1}$. If $\nu_{tp1} > \nu_{tn1}$, the restriction for neglect of one of the quadratic terms turns out to be that of (69) for the other one, but with subscripts n and p interchanged — an interchange that does not affect Δ_1. This distinction arises because the signs of the approximate linear terms in $\Delta\hat{n}$ depend on the sign of $\nu_{tn1} - \nu_{tp1}$. It is easily shown that† for this quantity zero, or the case of "critical recombination," $\Delta\hat{n}$ is identically zero and (67) are linear for all Δg. For trapping without recombination, (68) does not apply because Δ_1 is zero, but the conditions may properly be written as the restrictions on the concentrations obtained by use of these equations. For example, for electron trapping with C_{p1} zero, the condition $\Delta n \ll n_0 + n_1$ results, which may be a severe condition for p-type material.

† A solution is excluded that does not admit thermal equilibrium, for which Δg or $\Delta\hat{n}$ is zero for certain negative values of Δn and Δp or Δp and Δg, respectively.

3.1.2 *Nonlinear Theory*

For added carrier concentrations resulting from arbitrary injection levels, steady-state lifetime functions τ_n and τ_p may be evaluated from $\mathfrak{R}_m$. For acceptor centers only and no direct recombination, $\mathfrak{R}_m$ is given by

$$\begin{aligned}\mathfrak{R}_m &= C_{n1}[n(\mathfrak{N}_1 - \hat{n}) - n_1\hat{n}] \\ &= C_{p1}[p\hat{n} - p_1(\mathfrak{N}_1 - \hat{n})],\end{aligned} \tag{70}$$

which results in

$$\begin{aligned}\mathfrak{R}_m &= \frac{np - n_i^2}{\tau_{p0}(n + n_1) + \tau_{n0}(p + p_1)} \\ &= \frac{\hat{n}(\mathfrak{N}_1 - \hat{n}) + (n_0 + n_1)\hat{n} + (p_0 + p_1)(\mathfrak{N}_1 - \hat{n})}{\tau_{p0}(\mathfrak{N}_1 - \hat{n}) - \tau_{n0}\hat{n}} \frac{\Delta\hat{n}}{\mathfrak{N}_1} \\ &= \frac{\hat{n}(1 - \hat{n}/\mathfrak{N}_1) + (n_0 + n_1 - p_0 - p_1)\Delta\hat{n}/\mathfrak{N}_1 + n_0 + p_0}{\tau_{p0}(\mathfrak{N}_1 - \hat{n}) - \tau_{n0}\hat{n}} \Delta\hat{n} \\ &= \Delta n/\tau_n = \Delta p/\tau_p .\end{aligned} \tag{71}$$

Eliminating $\hat{n}$ by means of the second equation of (70) and the use of (61) results in the first form† for $\mathfrak{R}_m$ of (71). This familiar form[31] furnishes τ_n or τ_p in terms of Δn or Δp alone if one of these concentrations is eliminated by solving the second equation of (70) written with $\hat{n}$ replaced by $p - n$; and Δn or Δp may at the same time be related to, say, the generation rate $\Delta g = \mathfrak{R}_m$ for steady-state photoconductivity. The algebra involves radicals. A better procedure for such analysis employs the second or the third form for $\mathfrak{R}_m$; these result from (70) by elimination of n and p with the neutrality condition. Then Δn and Δp are, with $\mathfrak{R}_m$, written in term of $\hat{n}$ or $\Delta\hat{n}$ as independent parameter in accordance with‡

$$\Delta n = \frac{\tau_{p0}(n_0 + n_1) + \tau_{n0}(\hat{n} + p_0 + p_1)}{\tau_{p0}(\mathfrak{N}_1 - \hat{n}) - \tau_{n0}\hat{n}} \Delta\hat{n}$$

and (72)

$$\Delta p = \Delta m = \frac{\tau_{n0}(p_0 + p_1) + \tau_{p0}(\mathfrak{N}_1 - \hat{n} + n_0 + n_1)}{\tau_{p0}(\mathfrak{N}_1 - \hat{n}) - \tau_{n0}\hat{n}} \Delta\hat{n},$$

so that the lifetime functions are given by

† If two or more different kinds of centers are present, then $\mathfrak{R}_m$ is the sum of similar terms for each kind. For the corresponding lifetimes in terms of Δn and Δp, see Okada.[45]

‡ Note that $\Delta\hat{n}$ has the sign of the denominator, which is proportional to $\nu_{tn1} - \nu_{tp1}$ for $\Delta\hat{n}$ small.

$$\tau_n = \frac{\tau_{p0}(n_0 + n_1) + \tau_{n0}(\hat{n} + p_0 + p_1)}{\hat{n}(1 - \hat{n}/\mathfrak{N}_1) + (n_0 + n_1 - p_0 - p_1)\Delta\hat{n}/\mathfrak{N}_1 + n_0 + p_0}$$

and (73)

$$\tau_p = \frac{\tau_{n0}(p_0 + p_1) + \tau_{p0}(\mathfrak{N}_1 - \hat{n} + n_0 + n_1)}{n(1 - \hat{n}/\mathfrak{N}_1) + (n_0 + n_1 - p_0 - p_1)\Delta\hat{n}/\mathfrak{N}_1 + n_0 + p_0}.$$

The functions reduce to the Shockley-Read lifetimes of (60) for the equilibrium value $\hat{n}_0$ of $\hat{n}$, as may be verified by use of (41) and (42). The range of $\hat{n}$ is from $\hat{n}_0$ to the limiting large-signal value given by $\hat{n}/\mathfrak{N}_1 = \tau_{p0}/(\tau_{n0} + \tau_{p0})$, for which the denominator in (72) vanishes, and for which τ_n and τ_p both equal[31] $\tau_{n0} + \tau_{p0}$.

The trapping ratio r_n corresponding to the lifetime functions of (73) is given by

$$\begin{aligned} r_n^{-1} &= \frac{\tau_{p0}(\mathfrak{N}_1 - \hat{n} + n + n_1) + \tau_{n0}(p + p_1)}{\tau_{p0}(\mathfrak{N}_1 - \hat{n}) - \tau_{n0}\hat{n}} \\ &= \left[1 + \frac{\tau_{p0}\Delta n + \tau_{n0}\Delta p}{\tau_{p0}(\mathfrak{N}_1 - \hat{n} + n_0 + n_1) + \tau_{n0}(p_0 + p_1)}\right]\frac{\Delta p}{\Delta\hat{n}} \\ &\quad - \frac{\Delta p}{\Delta\hat{n}} + \frac{\begin{array}{r}\tau_{p0}^2(n_0 + n_1) + \tau_{n0}^2(p_0 + p_1) \\ + \tau_{p0}\tau_{n0}(\mathfrak{N}_1 + n_0 + n_1 + p_0 + p_1)\end{array}}{[\tau_{p0}(\mathfrak{N}_1 - \hat{n}) - \tau_{n0}\hat{n}]^2}\,\Delta\hat{n}, \end{aligned} \tag{74}$$

which is obtained from (72) by differentiating with respect to Δm. The equilibrium value of r_n, which is that of $\Delta\hat{n}/\Delta p$, is the r_n given in (60), while the limiting large-signal value is zero, as may be expected. By means of (72) and (74), the steady-state continuity equation, (30), may be written (for acceptor centers only), with $\Delta\hat{n}$ as independent variable and the components of grad $\Delta\hat{n}$ as dependent variables. The second or third form for r_n^{-1} of (74), with $\Delta p/\Delta\hat{n}$ given by (72), lends itself to this purpose; note that $d\Delta n/d\Delta\hat{n}$ and $d\Delta p/d\Delta\hat{n}$ equal $r_n^{-1} - 1$ and r_n^{-1}, respectively.

The lifetime function[31] for $|\Delta\hat{n}| \ll \Delta n \sim \Delta p = \Delta m$,

$$\tau_n \sim \tau_p \sim \frac{\tau_{p0}(n_0 + n_1 + \Delta p) + \tau_{n0}(p_0 + p_1 + \Delta p)}{n_0 + p_0 + \Delta p}, \tag{75}$$

may be derived most directly from the first form for $\mathfrak{R}_m$ of (71). By solving the second equation of (70) for $\Delta\hat{n}$, the condition $|\Delta\hat{n}| \ll \Delta n \sim \Delta p$ may be written as

$$\Delta p + \frac{\nu_{gn1} + \nu_{gp1}}{C_{n1} + C_{p1}} \gg \frac{|\nu_{tn1} - \nu_{tp1}|}{C_{n1} + C_{p1}} = \frac{|\nu_{gn1}/n_0 - \nu_{gp1}/p_0|}{C_{n1} + C_{p1}}\,\mathfrak{N}_1^*. \tag{76}$$

Equilibrium release probabilities are here employed. Equation (76) with Δp set equal to zero is the condition that $\Delta\hat{n}$ be relatively small for all Δp. Since this condition is (66) with ϵ set equal to zero, it subsumes the condition for equilibrium lifetimes substantially τ_0, to which it is usually equivalent.

It is readily shown from (75) that, with† $\Delta p \ll n_0 + p_0$ (and equilibrium lifetimes τ_0), the lifetimes are substantially τ_0 if Δp is small compared with $(\nu_{gn1} + \nu_{gp1})/(C_{n1} + C_{p1})$. This condition and the one of (66) may be severe conditions under essentially the same circumstances. That is, in the minority-carrier trapping range defined in Section 2.2.1.2, lifetimes are τ_0 for $\mathfrak{N}_1^*$ small compared with minority-carrier concentration n_0 or p_0; and then, consistent also with the condition of Section 3.1.1 for the neglect of $C_{n1}\Delta n\Delta\hat{n}$ or $C_{p1}\Delta p\Delta\hat{n}$ suitably specialized, for Δp small compared with $n_0 + n_1$ or $p_0 + p_1$. If the condition on Δp is not met, then, with the condition on $\mathfrak{N}_1^*$, (75) gives a lifetime that increases rapidly with injection level at low injection levels.‡ But such observed behavior with extrinsic material, as these considerations indicate, cannot usually be properly analyzed by use of (75). The steady-state lifetimes in the small-signal range generally either result primarily from recombination or majority-carrier trapping and are both τ_0 and substantially constant, or else have distinct equilibrium values given by (60) with dependences on (small-signal) injection level obtainable by the general procedure described. It will be shown that, in the latter case, substantially constant apparent diffusion-length lifetimes given by τ_{gn1} for n-type material or τ_{gp1} for p-type generally apply in the small-signal range above a certain injection level. Thus, unless trap concentration is quite small, (75) has significant application in the former case only to the transition from τ_0 to the lifetime $\tau_{n0} + \tau_{p0}$ for the large-signal range.

Nonconstant small-signal lifetime functions are associated with deep traps in the minority-carrier trapping range. Such traps will be saturated (in the steady state) even in the presence of a concentration of mobile minority carriers that is relatively quite small. From (67), by equating $d\Delta n/dt$ and $d\Delta p/dt$ (which are zero also in the immediate context), $\Delta\hat{n}$ may be written as

$$\Delta\hat{n} = (\nu_{tn1}\Delta n - \nu_{tp1}\Delta p)/(\nu_{gn1} + \nu_{gp1} + C_{n1}\Delta n + C_{p1}\Delta p), \quad (77)$$

† The more general condition without this restriction includes $\tau_0 \sim \tau_{n0} + \tau_{p0}$ for large values of Δp.

‡ As a result of saturation of centers available for minority-carrier capture, this lifetime increases essentially linearly in the small-signal range from the equilibrium value $\tau_{p0}(n_0 + n_1)/p_0$ or $\tau_{n0}(p_0 + p_1)/n_0$ and asymptotically to the large-signal value τ_{p0} or τ_{n0}.

in which the concentration-dependent contributions to the release frequencies are exhibited separately. Suppose that the transitions in p-type material are electron-dominated, so that $\nu_{gn1} \gg \nu_{gp1}$. For deep traps, as defined in Section 2.2.1.2, $n_1 \ll p_0$ (or $p_1 \gg n_0$) holds, which gives $n_0 + n_1 \ll p_0 + p_1$ and therefore implies $C_{n1} \ggg C_{p1}$ or $\tau_{p0} \ggg \tau_{n0}$ for the present case. Then, for $\Delta n \gg n_0 + n_1$, the denominator in (77) is $C_{n1}\Delta n$. Furthermore, $\nu_{tn1} \equiv C_{n1}(\mathfrak{N}_1 - \hat{n}_0) \ggg \nu_{tp1} \equiv C_{p1}\hat{n}_0$ holds from (53). Hence $\Delta\hat{n} \sim \mathfrak{N}_1 - \hat{n}_0$ follows† for Δn large compared with $n_0 + n_1$ and not too small compared with Δp. If trap concentration is not too large, small-signal saturation evidently occurs under the conditions assumed. If it is large, then a large conductivity increase is associated with the majority-carrier concentration corresponding to the saturated traps. Nonconstant small-signal lifetime functions apply in either case, whether saturation occurs in the small-signal range or not.

The lifetime functions in the saturation range approach the limiting large-signal lifetime, $\tau_{n0} + \tau_{p0}$, substantially equal to τ_{p0}. Though τ_{p0} is otherwise the minority-hole-capture-limited hole lifetime in strongly extrinsic n-type material, in this case it is a lifetime limited by majority-carrier capture. For small-signal saturation, $\Delta\hat{n}$ changes relatively slightly from the small-signal saturation range to its limiting value, $(\nu_{tn1} - \nu_{tp1})/(C_{n1} + C_{p1}) = \mathfrak{N}_1 - \hat{n}_0 - \tau_{n0}\mathfrak{N}_1/(\tau_{n0} + \tau_{p0})$, for the large-signal range. This circumstance might suggest that τ_{p0} applies over both ranges. In general, it does not: The denominators in (73) are comparatively small (reducing, for example, to $n_0 + n_1$ for $\Delta\hat{n} = \mathfrak{N}_1 - \hat{n}_0$) and are sensitive to very small changes in $\Delta\hat{n}$; a very small change in concentration of unsaturated traps can affect lifetimes appreciably. As will be shown in Section 3.2.2, large-signal lifetime implies relatively large increase in conductivity. The equations of (67) for the steady state, simplified for relatively small departure of $\Delta\hat{n}$ from $\mathfrak{N}_1 - \hat{n}_0$ still (necessarily) nonlinear, may be solved, in terms of Δn and for the saturation range, for the lifetime functions $\tau_n = \Delta n/\Delta g$ and $\tau_p = \Delta p/\Delta g$. With $\nu_{gn1} \gg \nu_{gp1} + \tau_{p0}^{-1}$, which will still apply in the present case even if $\mathfrak{N}_1$ is of order $p_0 + p_1$, it is found that $\tau_n \sim \tau_{p0}/[1 + (p_0 + \mathfrak{N}_1 - \hat{n}_0)/\Delta n]$ holds for the saturation range, specified by $\Delta n \gg n_0 + n_1$. Thus,

$$\tau_n \sim \tau_{p0}\Delta n/(p_0 + \mathfrak{N}_1 - \hat{n}_0),$$

proportional to Δn, holds for the saturation range of relatively small Δn. An apparent diffusion-length lifetime, τ_0', may be found by evaluating

† The equivalent condition $\Delta p \gg \mathfrak{N}_1 - \hat{n} + n_0 + n_1$ from (72) takes into account $\Delta p > \Delta n$. The small-signal saturation value of $\mathfrak{N}_1 - \hat{n}$ may be appreciably larger than $n_0 + n_1$, but its limiting large-signal value is small compared with $n_0 + n_1$ if $\mathfrak{N}_1 \ll (\nu_{gn1}/\nu_{gp1})(p_0 + p_1)$ holds.

$(D'/D)\tau_p = (1 - r_n)\tau_p$ for small-signal saturation.† With $r_n^{-1} = 1 + d\Delta n/d\Delta\hat{n}$, the expression $1 + (n_0 + n_1)(\mathfrak{N}_1 - \hat{n}_0)/\Delta n^2$ is found for $1 - r_n$; and dividing $\Delta p \sim \Delta n + \mathfrak{N}_1 - \hat{n}_0$ by Δg gives

$$\tau_p \sim [1 + (\mathfrak{N}_1 - \hat{n}_0)/\Delta n]\tau_n$$

$$\sim \tau_{p0}[1 + (\mathfrak{N}_1 - \hat{n}_0)/\Delta n]/[1 + (p_0 + \mathfrak{N}_1 - \hat{n}_0)/\Delta n].$$

It is easily seen that, for small $\mathfrak{N}_1 - \hat{n}_0$ of order $n_0 + n_1$ or less, the lifetime function for τ_0' so obtained is $\tau_{p0}\Delta n/p_0$, as are τ_n and τ_n . As may be expected, this result is consistent with (75). For $\mathfrak{N}_1 - \hat{n}_0 \ggg n_0 + n_1$, however, the lifetime function gives $\tau_0' \sim \tau_{p0}/(1 + \tau_{p0}/\tau_{gp1})$ for $(\mathfrak{N}_1 - \hat{n}_0)^2 \gg \Delta n^2 \gg (n_0 + n_1)(\mathfrak{N}_1 - \hat{n}_0)$. The condition $\mathfrak{N}_1 \ll p_0 + p_1$ then gives $\tau_0' \sim \tau_{gp1}$ for small-signal saturation. If this inequality is reversed, then $\tau_0' \sim \tau_{p0}$ results, and saturation occurs with relatively large increase in hole concentration.

From (60), the equilibrium electron, hole and diffusion-length lifetimes are, for these cases, generally small compared with τ_{p0} . They are given by $\tau_0 \sim \tau_n \sim [(n_0 + n_1)/p_0]\tau_{p0} = [(n_0 + n_1)/(\mathfrak{N}_1 - \hat{n}_0)]\tau_{gp1}$ and $\tau_p \sim [(\mathfrak{N}_1 - \hat{n}_0)/p_0]\tau_{p0} = \tau_{gp1}$ if small-signal saturation occurs, for which τ_0 is also small compared with $\tau_0' = \tau_{gp1}$. The minority-carrier and apparent diffusion-length lifetime functions increase with injection level, most rapidly as Δn becomes comparable with $n_0 + n_1$ and the traps fill. These results clearly provide a simple model, based on a single trapping level, for superlinearity,[29,34,35,36] the more-rapid-than-linear increase of photoconductivity with injection level.‡ With small-signal saturation, two superlinear ranges may occur, the first as diffusion-length lifetime increases from τ_0 to τ_{gp1} , and the second as it increases from τ_{gp1} to τ_{p0} in the large-signal range. With large-signal saturation resulting from large concentration of traps, only one superlinear range occurs, since a nearly linear intermediate range is absent. Only one range occurs also under the condition of (66) for small trap concentration. It is evident, however, that with superlinearity this condition is generally quite severe.

For the majority-carrier-dominated case of $\nu_{tp1} \gg \nu_{tn1}$ (or $\nu_{gp1} \ggg \nu_{gn1}$) in p-type material, there can be no small-signal saturation. With small trap concentration, lifetime $\tau_0 \sim (1 + p_1/p_0)\tau_{n0} = \tau_{tn1}$, which is limited

† The second form follows since $\Delta n/(n_0 + n_1) \gg (\mathfrak{N}_1 - \hat{n}_0)/(p_0 + \mathfrak{N}_1 - \hat{n}_0)$ holds *a fortiori*.

‡ For small concentration of centers, Δp may exhibit a less-rapid-than-linear, a linear, or a superlinear dependence on Δg, as Rittner[1] has shown using a lifetime function tantamount to that of (75). From this equation, superlinearity results, as may be expected, if $\tau_{n0} + \tau_{p0}$ exceeds τ_0 , so that the numerator increases more rapidly with Δp than the denominator. See also Ridout,[70] Newman, Woodbury and Tyler[71] and Sandiford.[7]

by minority-carrier capture and obtains over the entire small-signal range, then changes in accordance with (75) to $\tau_{n0} + \tau_{p0}$ in the large-signal range. This change is a decrease to $\tau_{n0} \gg \tau_{p0}$ if the trapping level is near the Fermi level or higher.

The steady-state fractions of ionized centers can be represented by simple formal generalizations of the equilibrium relationships of (41) and (56). In these equations, $\hat{n}_0$ and $\hat{p}_0$ are replaced by $\hat{n}$ and $\hat{p}$ and α_{10} and α_{20} by

$$\begin{aligned} \alpha_1 &\equiv \frac{C_{n1}\, n + C_{p1}p_1}{C_{p1}p + C_{n1}n_1} = \alpha_{10}\frac{1 + \Delta n/(n_0 + n_1^*)}{1 + \Delta p/(p_0 + p_1^*)}, \\ \alpha_2 &\equiv \frac{C_{p2}p + C_{n2}n_2}{C_{n2}n + C_{p2}p_2} = \alpha_{20}\frac{1 + \Delta p/(p_0 + p_2^*)}{1 + \Delta n/(n_0 + n_2^*)}, \end{aligned} \tag{78}$$

as can readily be shown† by solving for the ionized fractions from (37) and also from the corresponding two-level equations of Section 2.2.2.

3.2 *Photoconductivity*

A number of results for steady-state photoconductivity being implicit in Section 3.1, the present section will deal principally with the transient decay.

3.2.1 *Linear Theory*

For two types of centers in the linear small-signal case the time derivatives of Δm, $\Delta\hat{n}$ and $\Delta\hat{p}$ for photoconductive decay are given respectively by (46) without Δg and the quadratic terms. The general solution is accordingly

$$\begin{aligned} \Delta m &= \sum_{j=1}^{3} A_j e^{-\nu_j t}, \\ \Delta\hat{n} &= \sum_{j=1}^{3} r_{nj} A_j e^{-\nu_j t}, \\ \Delta\hat{p} &= \sum_{j=1}^{3} r_{pj} A_j e^{-\nu_j t}, \end{aligned} \tag{79}$$

in which the A_j are constants determined by the initial conditions, and the r_{nj} and r_{pj} are trapping ratios for the respective decay modes determined by

† The equation given in the abstract of the paper of Sah and Shockley[22] rewritten in the present notation yields $\hat{n}/(\mathfrak{N} - \hat{n} - \hat{p}) = \alpha_1$ and $(\mathfrak{N} - \hat{n} - \hat{p})/\hat{p} = \alpha_2^{-1}$, from which the ionized fractions for the two-level case here given follow as solutions of simultaneous linear equations.

$$\begin{aligned}(\nu_{11} + \nu_j) + \nu_{12}r_{nj} + \nu_{13}r_{pj} &= 0,\\ \nu_{21} + (\nu_{22} + \nu_j)r_{nj} + \nu_{23}r_{pj} &= 0,\\ \nu_{31} + \nu_{32}r_{nj} + (\nu_{33} + \nu_j)r_{pj} &= 0,\end{aligned} \tag{80}$$

with the decay constants ν_j being the roots of the equation obtained by equating to zero the determinant of (80). The A_j are found in terms of the trapping ratios and the initial concentrations Δm_1, $\Delta\hat{n}_1$ and $\Delta\hat{p}_1$ by setting t equal to zero in (79) and solving. The solution so obtained applies as well to the two-level case, for which the ν_{ij} are defined in accordance with (57). The decay constants ν_j, which are roots of a cubic, are always real and (since the coefficients alternate in sign) positive, except that one of them may be zero. Establishing these properties involves expressing the coefficients in terms of the capture and release frequencies by means of (51) or (57) and making use of (53).

The constant term of the cubic and one decay constant are zero if there is trapping only and no recombination. This case can occur essentially in two ways: The two types of center may trap, respectively, the two kinds of carriers, or they may both trap only one kind. If, say, the acceptor centers trap only electrons and the donor centers trap only holes, then the ν_j are readily found to be zero, $\nu_{tn1} + \nu_{gn1}$ and $\nu_{tp2} + \nu_{gp2}$. The last two decay constants characterize the respective exponential increases with time of $\Delta\hat{n}$ and $\Delta\hat{p}$ to new equilibrium concentrations after injection that correspond to the zero decay constant. These equilibrium concentrations are fractions of Δm (which remains constant) equal to the fractions of the time the electrons and the holes are trapped. Indeed, since the two types of centers trap independently in this case, the solution consists of solutions written independently for each. But if, for example, electrons only are trapped by both types of centers, then this independence does not obtain; electrons released from centers of one type may be trapped by centers of the other type. With the convention here employed, concentration of electrons trapped by donor centers may then be written as negative $\Delta\hat{p}$. The decay constants are found to be zero and $\frac{1}{2}\{\nu_{tn1} + \nu_{gn1} + \nu_{tn2} + \nu_{gn2} \pm [(\nu_{tn1} + \nu_{gn1} - \nu_{tn2} - \nu_{gn2})^2 + 4\nu_{tn1}\nu_{tn2}]^{\frac{1}{2}}\}$, with an equilibrium value after injection of total trapped electron concentration equal to $\Delta p/[1 + \nu_{gn1}\nu_{gn2}/(\nu_{tn1}\nu_{gn2} + \nu_{tn2}\nu_{gn1})]$.

The general linear small-signal case for one type of center is readily evaluated in detail. For acceptor centers only, the solution is given by the first two equations of (79), all terms with $j = 3$ being omitted. The trapping ratios are given by

$$\begin{aligned}r_{nj} &= -\nu_{21}/(\nu_{22} + \nu_j) = (\nu_{tn1} - \nu_{tp1})/(\nu_{tn1} + \nu_{gn1} + \nu_{gp1} - \nu_j)\\ &= -(\nu_{11} + \nu_j)/\nu_{12} = (\nu_j - \nu_{tp1})/\nu_{gp1},\end{aligned} \tag{81}$$

and the ν_j are the roots of

$$\nu_j^2 - \nu_s\nu_j + \Delta_1 = 0, \qquad j = 1, 2, \tag{82}$$

where ν_s is defined by

$$\nu_s \equiv -(\nu_{11} + \nu_{22}) = \nu_{tn1} + \nu_{gn1} + \nu_{tp1} + \nu_{gp1} \tag{83}$$

and Δ_1 by (62). The decay constants are thus

$$\nu_j = \tfrac{1}{2}[(-1)^{j-1}\nu_r + \nu_s], \qquad j = 1, 2, \tag{84}$$

with

$$\nu_r \equiv (\nu_s^2 - 4\Delta_1)^{\frac{1}{2}} = [(\nu_{tn1} + \nu_{gn1} - \nu_{tp1} - \nu_{gp1})^2 + 4\nu_{gn1}\nu_{gp1}]^{\frac{1}{2}}. \tag{85}$$

The corresponding time constants $\tau_1 \equiv \nu_1^{-1}$ and $\tau_2 \equiv \nu_2^{-1}$ are also equal respectively to ν_2/Δ_1 and ν_1/Δ_1. Nonoscillatory decay is easily verified for this case: The second form for ν_r shows that the ν_j are real; and, since $\nu_r < \nu_s$, the ν_j are positive.

A subcase that provides some physical interpretations is that of $\mathfrak{N}_1$ sufficiently small so that capture frequencies are small compared with release frequencies. As (62) and (83) show, the condition $\nu_s^2 \gg 4\Delta_1$ then holds, and expansion of the radical in (85) gives

$$\begin{aligned} \tau_1 \sim \nu_s^{-1} &= \tau_{gn1}\tau_{gp1}/(\tau_{gn1} + \tau_{gp1}) \\ \ll \tau_2 \sim \nu_s/\Delta_1 &\sim \tau_0 . \end{aligned} \tag{86}$$

Thus, for this subcase, τ_2 is the steady-state lifetime τ_0 of (65). It is large compared with τ_1, the time constant for the adjustment of $\Delta\hat{n}$ to a fixed fraction of Δp substantially equal to the equilibrium trapping ratio, r_n. This interpretation of τ_1 follows from solutions for the concentrations: The last form for the r_{nj} of (81) and r_n from (60) give

$$\begin{aligned} r_{n1} &\sim 1 + \nu_{gn1}/\nu_{gp1} , \\ r_{n2} &\sim (\nu_{tn1} - \nu_{tp1})/(\nu_{gn1} + \nu_{gp1}) \sim r_n ; \end{aligned} \tag{87}$$

and the result

$$\begin{aligned} \Delta n/\Delta p_1 &= [r_n\nu_{gn1}/(\nu_{gn1} + \nu_{gp1})]e^{-t/\tau_1} \\ &\quad + [1 - r_n\nu_{gn1}/(\nu_{gn1} + \nu_{gp1})]e^{-t/\tau_2}, \\ \Delta\hat{n}/\Delta p_1 &= r_n(-e^{-t/\tau_1} + e^{-t/\tau_2}), \\ \Delta p/\Delta p_1 &= -[r_n\nu_{gp1}/(\nu_{gn1} + \nu_{gp1})]e^{-t/\tau_1} \\ &\quad + [1 + r_n\nu_{gp1}/(\nu_{gn1} + \nu_{gp1})]e^{-t/\tau_2} \end{aligned} \tag{88}$$

holds for initial concentration $\Delta\hat{n}_1$ zero. Since r_n is small, the mobile-carrier concentrations are mostly in the second or lifetime decay mode and differ only slightly. For this subcase, if the initial trapping ratio is r_n rather than zero, then the first decay modes are not present and $\Delta\hat{n}$ is $r_n\Delta p = r_n\Delta p_1 e^{-t/\tau_2}$. The first decay modes do not occur either for "critical recombination," with which $\Delta\hat{n}$ remains identically zero as a result of equal capture frequencies ν_{tn1} and ν_{tp1} or, for this subcase, equal capture rates for Δn and Δp. For small $\mathfrak{N}_1$, the capture rates are in all cases substantially in the ratio ν_{tn1}/ν_{tp1}. In linear cases, they also decay in the lifetime mode after this mode predominates. The release rates behave similarly, their ratio being equal to ν_{gn1}/ν_{gp1}, or to $(n_0/p_0)(\nu_{tn1}/\nu_{tp1})$ in accordance with (53).

The condition for neglect of the capture frequencies may be severe. The approximate form of ν_s applies if $\nu_{tn1} + \nu_{tp1} \ll \nu_{gn1} + \nu_{gp1}$ holds, which implies that

$$\begin{aligned}\mathfrak{N}_1^* &\ll (\nu_{gn1} + \nu_{gp1})/(\nu_{gn1}/n_0 + \nu_{gp1}/p_0) \\ &= (n_0\nu_{tn1} + p_0\nu_{tp1})/(\nu_{tn1} + \nu_{tp1}),\end{aligned} \tag{89}$$

a condition which subsumes

$$\mathfrak{N}_1^* \ll n_0 + p_0 \tag{90}$$

for neglect of $\nu_{tn1}\nu_{tp1}$ in Δ_1. The conditions of (89) and that for steady-state lifetimes equal to τ_0 of (66) are the same for the minority-carrier trapping range defined in Section 2.2.1.2, for which they are $\mathfrak{N}_1^*$ small compared with the equilibrium minority-carrier concentration. The condition $\nu_s^2 \gg 4\Delta_1$ is

$$\begin{aligned}\mathfrak{N}_1^* &\ll \tfrac{1}{4}n_i^2(n_0 + p_0)^{-1}(\nu_{gn1}/\nu_{gp1} + \nu_{gp1}/\nu_{gn1} + 2) \\ &= \tfrac{1}{4}n_i^2(n_0 + p_0)^{-1}(\tau_{gn1} + \tau_{gp1})^2/\tau_{gn1}\tau_{gp1}\end{aligned} \tag{91}$$

if (89) holds, and it can be shown to be weaker than (89) in general if the minority- to majority-carrier release frequency ratio exceeds a number that is about three for extrinsic material and about six for intrinsic material.† Equations (89) and (91) are both subsumed *a fortiori* by $\mathfrak{N}_1^* \ll n_i^2/(n_0 + p_0)$, which is $\mathfrak{N}_1 \ll (n_0 + n_1)(p_0 + p_1)/(n_0 + p_0)$.

The release frequencies may be neglected under the condition of (90) but with the inequality signs reversed. The solution is then simply‡ $\Delta n/\Delta n_1 = e^{-t/\tau_{tn1}}$ and $\Delta p/\Delta p_1 = e^{-t/\tau_{tp1}}$. For $\mathfrak{N}_1^*$ large, (60) shows that

† Equation (91) gives a stronger or weaker condition according to whether $n_0\nu_{gp1}/\nu_{gn1} + p_0\nu_{gn1}/\nu_{gp1}$ is smaller or larger than $3(n_0 + p_0)$.

‡ This result easily follows directly from the differential equations. Or, note that the radicand in (85) is $(\nu_{tn1} - \nu_{tp1})^2$.

τ_{tn1} and τ_{tp1} are respectively the steady-state lifetimes τ_n and τ_p. The condition $\nu_s^2 \gg 4\Delta_1$ is accordingly $\frac{1}{4}(\tau_n/\tau_p + \tau_p/\tau_n) + \frac{1}{2} \gg 1$ — namely, that one of τ_n or τ_p be small compared with the other. If τ_n or τ_p is the smaller, then substantially all of Δn or Δp, respectively, is transformed comparatively rapidly into positive or negative $\Delta\hat{n}$, after which a slower recombinative decay of $\Delta\hat{n}$ and the concentration of the other mobile carriers takes place as these carriers are captured.

The condition $\nu_s^2 \gg 4\Delta_1$ implies $\tau_1 \ll \tau_2$, with τ_1 essentially a characteristic time for trapping and τ_2 essentially a lifetime. This interpretation does not apply if ν_s^2 and $4\Delta_1$ are comparable so that τ_1 and τ_2 do not differ by much. For small $\mathfrak{N}_1$ and the majority-carrier trapping range, for example, $\tau_1 \sim \tau_2$ may hold; (89) may apply, but not (91) (see footnote on previous page). The case of $\nu_s^2 \sim 4\Delta_1$ for $\mathfrak{N}_1$ large, for which τ_1, τ_2, τ_{tn1}, τ_{tp1}, τ_n and τ_p are all substantially equal, is a case of recombination with but slight trapping.

The general trapping time and lifetime obtained from (84) and related equations are

$$\begin{aligned}
\tau_1 = \nu_s^{-1} &= \tau_{gn1}\tau_{gp1}/[(\mathfrak{N}_1^*/p_0 + 1)\tau_{gn1} + (\mathfrak{N}_1^*/n_0 + 1)\tau_{gp1}] \\
&= \tau_{tn1}\tau_{tp1}/[(1 + p_0/\mathfrak{N}_1^*)\tau_{tn1} + (1 + n_0/\mathfrak{N}_1^*)\tau_{tp1}] \\
\ll \tau_2 = \bar{\nu}_s/\Delta_1 &= (\mathfrak{N}_1^* + n_0 + p_0)^{-1} \\
&\quad \cdot [n_0\tau_{gn1} + p_0\tau_{gp1} + n_i^2(\tau_{gn1} + \tau_{gp1})/\mathfrak{N}_1^*] \\
&= (\mathfrak{N}_1^* + n_0 + p_0)^{-1} \\
&\quad \cdot [(\mathfrak{N}_1^* + p_0)\tau_{tn1} + (\mathfrak{N}_1^* + n_0)\tau_{tp1}].
\end{aligned} \tag{92}$$

Comparison with (60) and (65) shows that this lifetime τ_2 is larger than the steady-state lifetimes τ_n, τ_p and τ_0; all are equal in the limit of $\mathfrak{N}_1$ small. For $\mathfrak{N}_1$ large in intrinsic material, τ_2 equals $2\tau_0$. Furthermore, these lifetimes all decrease monotonically to zero as $\mathfrak{N}_1$ increases indefinitely.

The decrease of τ_2 with increasing $\mathfrak{N}_1$ may, however, proceed essentially in two ranges, with approximate constancy of τ_2 in an intermediate range.[41] From the first form for τ_2 of (92), this intermediate range occurs provided there are capture concentrations $\mathfrak{N}_1^*$ that are small compared with $n_0 + p_0$ and also large compared with

$$(\nu_{gn1} + \nu_{gp1})/(\nu_{gn1}/n_0 + \nu_{gp1}/p_0);$$

that is, if the strong inequality

$$\tau_{gn1} + \tau_{gp1} \lll (n_0/p_0)\tau_{gn1} + (p_0/n_0)\tau_{gp1} \tag{93}$$

holds. It can hold for sufficiently strongly extrinsic material if the majority-carrier release time is not too small. For small $\mathfrak{N}_1$, τ_2 varies inversely with $\mathfrak{N}_1$, as (65) for τ_0 shows. For large $\mathfrak{N}_1$ such that $\mathfrak{N}_1^* \gg n_0 + p_0$, τ_2 varies similarly, equalling the value $(n_0\tau_{gn1} + p_0\tau_{gp1})/(n_0 + p_0)$ of approximate constancy divided by $\mathfrak{N}_1^*/(n_0 + p_0)$. With the third or fourth form for $\mathfrak{N}_1^*$ of (63), this τ_2 reduces to $\tau_{tn1} + \tau_{tp1}$. Since τ_1 for large $\mathfrak{N}_1^*$ is the harmonic mean of τ_{tn1} and τ_{tp1}, τ_1 is the smaller of these capture times and τ_2 the larger, as previously discussed for this case. It can be shown that, for the minority-carrier trapping range, the inequalities that $\mathfrak{N}_1^*$ must satisfy for approximate constancy of τ_2 generally imply the condition $\nu_s^2 \gg 4\Delta_1$ on which the calculation is based.† A similar situation has been shown to obtain with the inequality for the case of negligible capture frequencies. But since this case involves a condition for neglect of the capture frequencies that is usually severe in the minority-carrier trapping range, it is the present case that would usually apply in practice in this range.

3.2.2 *Nonlinear Theory*

Although the general problem of photoconductive decay is intractable analytically, some special cases can be solved and certain techniques of approximation are effective. From (46) and (51), general equations that apply for centers of the acceptor type may be written as

$$\begin{aligned} d\Delta n/dt &= \Delta g + \nu_{gn1}\Delta\hat{n} - [\nu_{tn1} - (C_{n1} - C_i)\Delta\hat{n}]\Delta n - C_i\Delta n^2, \\ d\Delta\hat{n}/dt &= [\nu_{tn1} - \nu_{tp1} - (C_{n1} + C_{p1})\Delta\hat{n}]\Delta n \\ &\quad - (\nu_{tp1} + \nu_{gn1} + \nu_{gp1})\Delta\hat{n} - C_{p1}\Delta\hat{n}^2 \\ &= [\nu_{tn1} - \nu_{tp1} - (C_{n1} + C_{p1})\Delta\hat{n}]\Delta p \\ &\quad - (\nu_{tn1} + \nu_{gn1} + \nu_{gp1})\Delta\hat{n} + C_{n1}\Delta\hat{n}^2, \\ d\Delta p/dt &= \Delta g - \nu_{gp1}\Delta\hat{n} - [\nu_{tp1} + (C_{p1} - C_i)\Delta\hat{n}]\Delta p - C_i\Delta p^2, \end{aligned} \tag{94}$$

in which equilibrium values of release frequencies are employed. Since $|\Delta\hat{n}|$ is bounded (by a concentration that cannot exceed $\mathfrak{N}_1$), it is clear that, if the initial concentration Δp_1 is sufficiently large, then the decay proceeds with $\Delta\hat{n}$ after a short transient substantially equal to

$$(\nu_{tn1} - \nu_{tp1})/(C_{n1} + C_{p1}),$$

† One inequality is the reverse of that of (89); hence it is the condition for the neglect of the release frequencies in ν_s. The other inequality is (90). It follows that $\nu_s^2 \gg 4\Delta_1$ is $\mathfrak{N}_1^* \gg 4(n_0^{-1} + p_0^{-1})\nu_{gn1}\nu_{gp1}/(\nu_{gn1}/n_0 + \nu_{gp1}/p_0)^2$. This condition is weaker than the reverse of (89) under the same circumstances that make (91) a weaker condition than (89).

for which the coefficient of Δn or Δp in $d\Delta\hat{n}/dt$ is zero. This value of $\Delta\hat{n}$ corresponds to maximum (but not necessarily complete) saturation or neutralization in the traps. Although direct recombination, characterized by the quasi-hyperbolic decay law

$$\Delta p/\Delta p_1 = (C_i \Delta p_1 t + 1)^{-1} \tag{95}$$

predominates in principle for very large Δp_1, it can frequently be neglected.[48,49,50] The large-signal decay is then exponential with lifetime $(C_{n1} + C_{p1})/(C_{n1}\nu_{tp1} + C_{p1}\nu_{tn1}) = \tau_{n0} + \tau_{p0}$, the limiting large-signal steady-state lifetime. With the limiting value of $\Delta\hat{n}$, this result follows from, say, the last equation of (94), in which neglect of the release rate $\nu_{gp1}\Delta(\mathfrak{N}_1 - \hat{n})$ is consistent with Δp large.

Examination in further detail of the large- and small-signal decay is facilitated by the equation for negligible direct recombination,

$$\tau_{n0}\Delta n^{-1} d\Delta n/dt + \tau_{p0}\Delta p^{-1} d\Delta p/dt + 1 + [(p_0 + p_1)(1 - \Delta n/\Delta p) + (n_0 + n_1)(1 - \Delta p/\Delta n)]/\mathfrak{N}_1 = (\tau_{n0}\Delta n^{-1} + \tau_{p0}\Delta p^{-1})\Delta g, \tag{96}$$

which is readily obtained from (67) or (94) as a linear combination of $d\Delta n/dt = d(\Delta p - \Delta\hat{n})/dt$ and $d\Delta p/dt$ that eliminates the quadratic terms. For example, consistent with results for the linear case of small $\mathfrak{N}_1$, assuming in this equation the steady-state trapping ratio r_n of (87) and (so that Δn and Δp are proportional) a single decay time, this time is given as the lifetime $\tau_2 \sim \tau_0$ of (86). If either $\Delta n \sim \Delta p$ or $\mathfrak{N}_1$ is sufficiently large, then the term with brackets, which arises from the terms involving release frequencies, may evidently be neglected. It is otherwise plausible that release does not appreciably affect large mobile carrier concentrations, while capture predominates with large trap concentration. For no volume generation, (96) may then be integrated, with the result

$$(\Delta n \Delta p)^{\frac{1}{2}}(\Delta n/\Delta p)^{-\frac{1}{2}(C_{n1} - C_{p1})/(C_{n1} + C_{p1})} = A\, e^{[-t/(\tau_{n0} + \tau_{p0})]}. \tag{97}$$

in which A is a constant determined by the initial concentrations. It is easily verified that, besides furnishing the large-signal lifetime, (97) is consistent with the linear solution for large $\mathfrak{N}_1$, for which the release frequencies may be neglected and the decay times are τ_{tn1} and τ_{tp1}. This equation is a first integral of (94) for a case of large $\mathfrak{N}_1$, one which can accordingly be formulated as a first-order (rather than second-order) nonlinear differential equation.

The condition under which (97) holds is, from (96),

$$\mathfrak{R}_1 \gg | \Delta\hat{n}[(n_0 + n_1)/\Delta n - (p_0 + p_1)/\Delta p] | . \tag{98}$$

From (96) and (98), the large-signal lifetime $\tau_{n0} + \tau_{p0}$ obtains if $\Delta n \sim \Delta p \gg \Delta\hat{n}$ holds and $\Delta p \gg \mathfrak{R}_1^{-1} | \Delta\hat{n}(n_0 + n_1 - p_0 - p_1) |$ also. With the $\Delta\hat{n}$ for $\Delta n \sim \Delta p$ from (77), these conditions are respectively that of (76) and

$$\begin{aligned} \Delta p + (\nu_{gn1} + \nu_{gp1})/(C_{n1} + C_{p1}) \\ = \Delta p + \tau_0(n_0 + p_0)/(\tau_{n0} + \tau_{p0}) \\ \gg | n_0 + p_0 - (\nu_{gn1} + \nu_{gp1})/(C_{n1} + C_{p1}) | \\ = (n_0 + p_0) | 1 - \tau_0/(\tau_{n0} + \tau_{p0}) | . \end{aligned} \tag{99}$$

Since the left-hand sides of (76) and (99) are the same, comparison of the right-hand sides will indicate which condition is the more restrictive in any particular case.† The condition that corresponds to (99) obtained from the lifetime for $\Delta n \sim \Delta p$ of (75) is similar except that the constant term on the left is replaced by $n_0 + p_0$. Setting Δp equal to zero in either gives the condition that $\tau_{n0} + \tau_{p0}$ apply for all Δp, which is that it equal τ_0.

The decay times associated with a small-amplitude pulse of added carriers above a steady generation level Δg are readily evaluated. The equations for $d\delta n/dt$ and $d\delta p/dt$, linear in the concentration increments δn and δp that result from the pulse, may be obtained from (67). Written with capture and release frequencies that are concentration-dependent, they are formally the same as the linear small-signal ones for $d\Delta n/dt$ and $d\Delta p/dt$. For the release frequencies, the definitions of (50) apply; for the capture frequencies, $\hat{n}_0$ in these definitions is replaced by $\hat{n}$. The condition $\nu_s^2 \gg 4\Delta_1$ of Section 3.2.1 generalized in this way is the condition for a lifetime $\tilde{\tau}_2$ for δn and δp equal to the generalized ratio ν_s/Δ_1 and large compared with the corresponding time constant for trapping. The lifetime $\tilde{\tau}_2$ depends on the steady-state values of Δn, $\Delta\hat{n}$ and Δp; (72) gives Δn and Δp in terms of $\Delta\hat{n}$, and (71) relates $\Delta\hat{n}$ to $\Delta g = \mathfrak{R}_m$. It reduces to τ_2 of (92) for the linear small-signal case and to $\tau_{n0} + \tau_{p0}$ for Δg large.

The approximation $\Delta n \sim \Delta p$ applied to (94) gives the differential equation

$$d\Delta p/dt + \frac{[1 + (n_0 + p_0)^{-1}\Delta p]\Delta p}{[1 + (\tau_{n0} + \tau_{p0})\tau_0^{-1}(n_0 + p_0)^{-1}\Delta p]\tau_0} = 0, \tag{100}$$

† In the minority-carrier trapping range, for example, (76) and (99) may require approximately that Δp be large compared with trap concentration and with equilibrium majority-carrier concentration, respectively.

whose solution[1,44] may be written as

$$\frac{\Delta p}{\Delta p_1}\left[\frac{1 + (n_0 + p_0)^{-1}\Delta p}{1 + (n_0 + p_0)^{-1}\Delta p_1}\right]^{[(\tau_{n0}+\tau_{p0})/\tau_0]-1} = e^{-(t/\tau_0)}. \tag{101}$$

As may be expected, this solution is the integrated form that corresponds to the lifetime of (75); the approximation results in the steady-state $\Delta\hat{n}$ of (77), and the condition required is that of (76). Casual inspection of (101) might suggest that τ_0 and $\tau_{n0} + \tau_{p0}$ apply respectively for Δp small and large compared with $n_0 + p_0$. This conclusion is, of course, illusory: For the minority-carrier trapping range, the exponent in (101) is large and τ_0 applies only if Δp is restricted as explained in Section 3.1.2.

The limiting large-signal $\Delta\hat{n}$, $\Delta\hat{p}$ and lifetime for two kinds of centers are easily evaluated from (46) and (51), and the extension to any number of kinds of centers is obvious. As may be expected, the values $(\nu_{tn1} - \nu_{tp1})/(C_{n1} + C_{p1})$ or $(\nu_{tp2} - \nu_{tn2})/(C_{n2} + C_{p2})$ of $\Delta\hat{n}$ or $\Delta\hat{p}$ are as if the acceptor or donor centers alone were present; and the lifetime is the harmonic mean of lifetimes $\tau_{n0} + \tau_{p0}$ for each kind of center, the decay constant being the sum of the separate decay constants. This result does not apply to the two-level case: From (57), $\Delta\hat{n}$ and $\Delta\hat{p}$ are found to equal $C_{n1}C_{n2}\mathfrak{N}/(C_{n1}C_{n2} + C_{n2}C_{p1} + C_{p1}C_{p2}) - \hat{n}_0$ and $C_{p1}C_{p2}\mathfrak{N}/(C_{n1}C_{n2} + C_{n2}C_{p1} + C_{p1}C_{p2}) - \hat{p}_0$, with $(1 + C_{n1}/C_{p1} + C_{p2}/C_{n2})/(C_{n1} + C_{p2})\mathfrak{N}$ as the large-signal lifetime.

General solutions for trapping only and no recombination can be obtained without difficulty. For, say, electron trapping by acceptor centers, Δp maintains its initial value Δp_1, and the nonlinear equation for Δn that results from replacing $\Delta\hat{n}$ by $\Delta p_1 - \Delta n$ in the first equation of (94) has the solution

$$\Delta n = \Delta n_2 \frac{1 - \dfrac{\nu_t + \nu_{tn1} - \nu_{gn1} - C_{n1}\Delta p_1}{-\nu_t + \nu_{tn1} - \nu_{gn1} - C_{n1}\Delta p_1} e^{-\nu_t t}}{1 - \dfrac{-\nu_t + \nu_{tn1} + \nu_{gn1} + C_{n1}\Delta p_1}{\nu_t + \nu_{tn1} + \nu_{gn1} + C_{n1}\Delta p_1} e^{-\nu_t t}} \tag{102}$$

for initial value Δp_1, if direct recombination is neglected,† with

$$\begin{aligned} \nu_t &\equiv [(\nu_{tn1} - \nu_{gn1} - C_{n1}\Delta p_1)^2 + 4\nu_{tn1}\nu_{gn1}]^{\frac{1}{2}}, \\ \Delta n_2 &\equiv \tfrac{1}{2}C_{n1}^{-1}(\nu_t - \nu_{tn1} - \nu_{gn1} + C_{n1}\Delta p_1). \end{aligned} \tag{103}$$

The concentration Δn_2 is the new equilibrium concentration which Δn

† The general form of this solution is not changed if direct recombination is included.

approaches asymptotically after injection. For Δp_1 large, these equations give

$$\Delta n = \left[\Delta p_1 - \left(1 - \frac{n_0 + n_1}{\Delta p_1}\right)(\mathfrak{N}_1 - \hat{n}_0)\right] \cdot \left(1 + \frac{\mathfrak{N}_1 - \hat{n}_0}{\Delta p_1} e^{-c_{n1}\Delta p_1 t}\right). \tag{104}$$

Thus, $\Delta\hat{n} = \Delta p_1 - \Delta n$ rapidly increases to the limiting value

$$[1 - (n_0 + n_1)/\Delta p_1](\mathfrak{N}_1 - \hat{n}_0) \sim \mathfrak{N}_1 - \hat{n}_0 = \nu_{tn1}/C_{n1},$$

which corresponds to substantially all traps charged. For Δp_1 small, the equations give

$$\Delta n/\Delta p_1 = [\nu_{gn1}/(\nu_{tn1} + \nu_{gn1})]\,[1 + (\nu_{tn1}/\nu_{gn1})\, e^{-(\nu_{tn1} + \nu_{gn1})t}], \tag{105}$$

which may be obtained also by suitable specialization of results for the general linear small-signal case. According to (105), $\Delta n/\Delta p_1$ decreases from unity to $\tau_{tn1}/(\tau_{tn1} + \tau_{gn1})$, the fraction of the time electrons are free, while $\Delta\hat{n}/\Delta p_1$ increases to $\tau_{gn1}/(\tau_{tn1} + \tau_{gn1})$, the fraction of the time electrons are trapped. An effect of slight recombination on Δn would in all cases be a comparatively slow decay from a value approximately equal to the equilibrium value Δn_2 for trapping only.

It is sometimes relevant to deal with a model involving centers that provide nonrecombinative trapping in conjunction with other centers, of a suitably idealized type, that provide only recombination that can be specified simply in terms of a constant lifetime. Such centers would in general be present in comparatively small concentration, so that the amplitude of their trapping transient is negligible. Furthermore, this transient would be comparatively brief, so that steady-state lifetime applies after negligible time.

With certain restrictions, the idealized centers may be centers that function in the recombination range or in the majority-carrier trapping range.† The $\Delta\hat{n}$ or $\Delta\hat{p}$ for these centers obtained by setting $d\Delta\hat{n}/dt$ or $d\Delta\hat{p}/dt$ equal to zero results in a contribution to both $d\Delta n/dt$ and $d\Delta p/dt$ that is the negative of a steady-state recombination rate similar to that of the first form for $\mathfrak{R}_m$ of (71). With subscripts "3" employed to denote the recombination centers, this recombination rate may be written as $\nu_{tn3}\Delta n/(1 + \nu_{gn3}/\nu_{gp3}) + \nu_{tp3}\Delta p/(1 + \nu_{gp3}/\nu_{gn3})$, in which the release frequencies are concentration-dependent. If now $\nu_{gp3} \gg \nu_{gn3}$ holds in p-

† For the minority-carrier trapping range, the lifetime function $(\tau_{n0} + \tau_{p0})\Delta p/(n_0 + p_0)$ would apply for $\Delta n \sim \Delta p \ll n_0 + p_0$, as can be shown by use of (100).

type material, it may be a consequence of the condition $\nu_{tp3} > \nu_{tn3}$ for hole trapping; or, with $\nu_{tp3} < \nu_{tn3}$, it may imply the recombination range. In either case, from (53) the small signal recombination rate is substantially $\nu_{tn3}\Delta n$ provided Δn is not too small because of strong electron trapping in the trapping centers. Lifetime τ_n is then $\tau_{tn3} \sim \tau_0$. While the large-signal lifetime differs in principle, the assumed inequality implies $\tau_0 \sim \tau_{n0} \gg \tau_{p0}$ if the energy level of the recombination centers is not too far from the Fermi level towards the valence band. In general, $\nu_{gp3} \gg \nu_{gn3}$ gives a small-signal recombination rate equal to

$$C_{n3}\mathfrak{N}_3(p_0\Delta n + n_0\Delta p)/(p_0 + p_3)$$

and thus a lifetime that cannot properly be associated with either Δn or Δp alone. In intrinsic material, for example, it is $\Delta n + \Delta p$ with which a lifetime may be associated.

These considerations suggest the formal representation of "linear recombination" by including in $\partial\Delta m/\partial t$ or in both $\partial\Delta n/\partial t$ and $\partial\Delta p/\partial t$ the negative of a recombination rate $\nu_{n3}\Delta n + \nu_{p3}\Delta p$. This procedure is useful in deriving results in forms that apply symmetrically without reference to conductivity type. For the p-type case here discussed, ν_{n3} and ν_{p3} equal ν_{tn3} and $(n_0/p_0)\nu_{tn3}$. The one that corresponds to the majority carrier can usually be set equal to zero for sufficiently strong extrinsic material.

3.2.3 *Negative Photoconductivity*

Under certain conditions, optical generation with excitations involving trapping levels will cause a decrease in conductivity below the thermal-equilibrium value.[51,52] This negative photoconductivity will be considered for a simple model — that of two types of centers, of which one gives trapping and the other only recombination. For traps of the acceptor type, (94) gives $d\Delta n/dt$ and $d\Delta p/dt$, except that suitable generation and recombination terms must be included. From Section 2.2.3, generation terms are respectively Δg_n and Δg_p ; and, from Section 3.2.2, the linear recombination term $-(\nu_{n3}\Delta n + \nu_{p3}\Delta p)$ may be included in both equations. For simplicity, direct recombination and the quadratic terms will be neglected, and the concentrations evaluated for the steady state. The result is

$$\Delta p = \frac{(\nu_{gp1} - \nu_{n3})\Delta g_n + (\nu_{tn1} + \nu_{gn1} + \nu_{n3})\Delta g_p}{\Delta_1 + (\nu_{n3} + \nu_{p3})(\nu_{gn1} + \nu_{gp1}) + \nu_{n3}\nu_{tp1} + \nu_{p3}\nu_{tn1}}, \quad (106)$$

with Δ_1 defined by (62); a similar expression for Δn is obtainable by

interchanging subscripts n and p, a transformation that does not change the denominator.

This result verifies the conclusion that Δp may be negative as a result of excitations from the traps to the conduction band in conjunction with recombination, and similarly for Δn, with excitations from the valence band to the traps. As a simple case, consider p-type material with ν_{p3} zero. If there is also trapping only of electrons and excitation only of electrons from traps to the conduction band, then C_{p1} and Δg_p are zero, whence Δn is zero and Δp is $-\Delta g_n/\nu_{gn1}$, with Δg_n equal to Δg_{c1} from (58). Recombination, in this case, produces negative Δp, which compensates the reduction by the excitation of the concentration of (negatively) charged traps, and the effect tends to be enhanced with deep traps of small capture cross section.

3.2.4 *Further Theory with an Application to Experiment*

Illustrative application will be made to observations of Hornbeck and Haynes on electron trapping in p-type silicon.[54] In this work, techniques were devised to measure the various time constants in the decay of photoconductivity, which, for certain samples, covered a range of about 10^7 in relative value. Evidence for two trapping levels was found, and electron capture cross sections and energy levels were estimated from the data, the model employed being that of two types of traps that capture only electrons, a lifetime being associated with recombination in centers of another type. The sample† for which there is most detailed information exhibited a 20-microsecond photoconductive decay, attributed to recombination, for sufficiently high injection levels; a decay of time constant about 10 milliseconds, attributed to decay in comparatively shallow traps that were initially filled in concentration of 2×10^{12} cm^{-3}; and decay in deep traps that were initially filled in concentration of 10^{13} cm^{-3} whose time constant varied from 1 second for the traps nearly full to 260 seconds for the traps nearly empty. Both types of traps are "deep" traps, as defined in Section 2.2.1.2. The present theory will be used to calculate the upper limits for the hole-capture cross sections implied by this model, and it will be shown how the conclusions are modified if an alternative model is assumed.

In outline, the general procedure here employed involves first assigning trial values to the energy levels of the traps, and then calculating expressions for decay constants from the equations, suitably linearized for particular ranges. These decay constants are roots of algebraic equa-

† Data and results for sample 223B are given in the text and various figures of the Hornbeck and Haynes paper.[54]

tions and, assuming them well separated, may be obtained as the magnitudes of ratios of successive coefficients. The coefficients are homogeneous expressions in capture and release frequencies (and also "constraint frequencies," if multiple-level traps are involved) — that of the highest power being unity, followed by linear, quadratic and higher-order forms for the successively lower powers. With assumed trapping levels and known equilibrium concentrations of carriers and unoccupied traps, the coefficients provide corresponding homogeneous forms in the capture coefficients. These coefficients remain to be found from observed decay constants. To each product of capture coefficients that occurs, a number of products of frequencies generally contribute, but, for the particular semiconductor material and trapping model, these usually differ by orders of magnitude and a single one predominates. With this considerable simplification, decay constants can be expressed in terms of the frequencies so that physical mechanisms involved can often be readily identified. If a sufficient number of distinct decay constants are known from experiment, the energies of the trapping levels may also be determined. The consistency of the assumed trapping model may then be checked; the energies found should clearly not differ from those assumed by so much that the particular simplified expressions employed for the decay constants do not apply.

Consistent with the notation here employed, the deep traps may be assumed to be of the donor type and the shallower traps of the acceptor type. Trial values of the energy levels will be taken as 0.23 ev below the Fermi level $\mathcal{E}$ in intrinsic material and at $\mathcal{E}$. These levels are approximately 0.78 ev below the conduction band for an energy gap of 1.10 ev and at midgap, substantially in accord with the locations determined by Hornbeck and Haynes† The values[72,73] at 300°K of 1500 and 570 cm^2 $volt^{-1}$ sec^{-1} for the electron and hole mobilities and 1.73×10^{20} cm^{-6} for n_i^2 give $n_0 = 4.3 \times 10^5$ cm^{-3} and $p_0 = 4.1 \times 10^{14}$ cm^{-3} for the 27-ohm-cm p-type sample, with $n_1 = p_1 = 1.32 \times 10^{10}$ cm^{-3}, $n_2 = 2.8 \times 10^6$ cm^{-3} and $p_2 = 6.1 \times 10^{13}$ cm^{-3}.

For the two kinds of traps with recombination at the rate $\nu_{n3}\Delta n \equiv \Delta n/\tau_3$ only in other centers of the idealized type discussed in Section 3.2.2, the outlined procedure applied to the equations written for the linear small-signal range gives the longest decay time τ_∞ as

$$\tau_\infty = \tau_3 + \tau_{gn1} + \tau_{gn2} + (\tau_{gn1}/\tau_{tn1})\tau_3 + (\tau_{gn2}/\tau_{tn2})\tau_3 . \quad (107)$$

The fourth and fifth terms represent recombination with multiple trap-

† An energy gap of 1.10 ev at 300°K is employed rather than 1.00 ev as in Ref. 54. The trial values employed originated in a two-level analysis (which later appeared inapplicable), the n_2 being that for $\mathfrak{N} = 1.15 \times 10^{13}$, or $\hat{p}_0/\mathfrak{N} = 0.87$.

ping in the shallower and deep traps, respectively, and the latter predominates for the case under examination.† In (107), τ_{gn1} and τ_{gn2} are, of course, the "effective" equilibrium release times, and are not correctly interpreted as the physically proper ones,‡ namely $(C_{n1}n_1)^{-1}$ and $(C_{n2}n_2)^{-1}$.

An upper limit for the coefficient C_{p2} for the capture of holes by occupied deep traps may be obtained by assuming recombination in deep traps only and then calculating τ_∞ from the linear small-signal equations. The result is

$$\tau_\infty = \frac{\nu_{tn2} + \nu_{gp2}}{\nu_{tn2}\nu_{gp2}} = \frac{\hat{p}_0 + \nu_{gp2}/C_{n2}}{\hat{p}_0\nu_{gp2}} \sim \tau_{gp2} \tag{108}$$

in the simplified form obtained by the outlined procedure. The final approximation on the right§ applies with $\hat{p}_0 = 10^{13}$ cm^{-3} and $\tau_\infty = 260$ seconds, provided merely that $C_{n2} \gg 4 \times 10^{-16}$ cm^3 seconds^{-1} holds. Then $C_{p2} = 8 \times 10^{-18}$ cm^3 seconds^{-1} follows, since $p_0 + p_2$ is 4.7×10^{14} cm^{-3}; and the cross section for hole capture A_{p2}, obtained by dividing by mean thermal velocity, is 8×10^{-25} cm^2. Recombination in the deep traps that gives τ_∞ cannot account for the observed decay. It can be shown‖ that the decay time for the traps nearly full would then be large compared with $\tau_{gp2} = 260$ seconds rather than 1 second. The actual A_{p2} may thus be considered small compared with about 10^{-24} cm^2.

For an upper limit to C_{p1}, recombination only in the shallower traps is assumed. For this case, the rather lengthy general expression for τ_∞ simplifies to give

$$\tau_\infty = \tau_{gn2} + (\tau_{gn2}/\tau_{tn2})[\tau_{tn1}(1 + \tau_{gp1}/\tau_{gn1})]. \tag{109}$$

The contribution τ_{gn2} is the time constant for the initial decay in the deep traps, obtainable as the longest decay time from the equations linearized for nearly full deep traps and nearly empty shallower ones. This release time represents recombination of electrons in the shallower

† See footnote 20 of Ref. 54.

‡ The τ_g of Ref. 54 should be identified as τ_{gn1} and τ_{gn2}. For the deep traps, $S\tau_g$ is accordingly $(n_0 + n_2)^{-1}$, which increases with the p-type conductivity and is not a property of the traps alone. The formula employed for locating trapping levels relative to a band edge holds if τ_g in it is the physically proper release time. With τ_g the "effective" release time, it holds only if n_0 is negligible compared with n_1 or n_2. In Equations (1) of Ref. 54, dn/dt lacks the term $C_{n1}n_0\Delta_n$.

§ The equilibrium τ_p of (60) and the lifetime τ_2 of (92) written for the deep traps, for which $\mathfrak{N}_2^* \sim 0.1\mathfrak{N}_2 \sim 0.1\hat{p}_0$, also reduce to τ_{gp2} if C_{n2} is not too small.

‖ The lifetime $\tilde{\tau}_2$ of Section 3.2.2 evaluated for the deep traps nearly saturated is, by use of approximations for near-saturation of Section 3.1.2, found to be given by $\tilde{\tau}_2 \sim (C_{p2}\mathfrak{N}_2)^{-1}\Delta n^2/(n_0 + n_2)p_0 \gg \tau_{gp2}$ for $(n_0 + n_2)p_0 \gg \Delta n^2 \gg (n_0 + n_2)\,\hat{p}_0$. The inequality on the right is equivalent to $\Delta n \gg \hat{p}$; the one on the left is largely consistent numerically with $\Delta n \ll n_0 + n_1$, for which the shallower traps are substantially empty.

traps without their recapture in the deep ones. The contribution $(\tau_{gn2}/\tau_{tn2})\tau_{tn1}$ represents recombination in the shallower traps with multiple trapping in the deep traps. The contribution with τ_{gp1}/τ_{gn1} as a factor represents recombination in the shallower traps with multiple trapping involving both levels.† This contribution predominates since, as shown in Section 3.1.2, the small-signal saturation of the shallower traps implies $\tau_{gp1} \gg \tau_{gn1}$. The only capture coefficient it contains is C_{p1}, which is found to equal 2.5×10^{-13} cm^3 seconds^{-1} for $\tau_\infty = 260$ seconds. The corresponding hole-capture cross section A_{p1} is 2.4×10^{-20} cm^2. The actual cross section is small compared with this value if recombination occurs primarily in centers of a third kind.

Recombination in the shallower traps can account for the observed deep-trap decay. Indeed, as may be expected, if the equations are linearized for small departures from a concentration $\hat{p}_1$ of unoccupied deep traps, then a (longest) decay time

$$\begin{aligned} \tau_d &= \tau_{gn2} + (\tau_{gn2}/\tau_{tn2})[\tau_{tn1}(1 + \tau_{gp1}/\tau_{gn1})](\hat{p}_1/\hat{p}_0) \\ &= \tau_{gn2} + (\tau_\infty - \tau_{gn2})(\hat{p}_1/\hat{p}_0) \end{aligned} \tag{110}$$

results, which increases from τ_{gn2} to τ_∞ as $\hat{p}_1$ increases from zero to $\hat{p}_0$ and is of the same form as that employed by Hornbeck and Haynes[54] to fit their data.‡ The observed decay in the shallower traps can also be accounted for through C_{p1}. The equations for nearly full deep traps and nearly empty shallower ones give $\tau_{gp1}(1 + \tau_{tn1}/\tau_{gn1}) \sim \tau_{gp1}$ for decay in the shallower traps as intermediate time constant, τ_{tn1} for electron capture being the shortest and τ_{gn2} for the initial deep-trap decay being the longest.§ The C_{p1} obtained by setting τ_{gp1} equal to 10^{-2} second is 3 per cent smaller than the value obtained from τ_∞ and is thus in rather fortuitously close agreement.

If a model with this C_{p1} is to account for experiment, then the assumption that the shallower traps are two-level traps, which gave the observed lifetime‖ of 20 microseconds through recombination in the higher

† The quantity in brackets in (109) can be shown, from (53), to be the τ_0 of (65) for the shallower traps in the p-type material. Thus, τ_0 itself may be said to entail multiple trapping through $(\tau_{gp1}/\tau_{gn1})\tau_{tn1}$, the major contribution to τ_0 in the minority-carrier trapping range.

‡ The interpretation differs, since τ_{gn2} is an "effective" release time. In the notation of Hornbeck and Haynes, $\hat{p}_1/\hat{p}_0$ is $1 - y$.

§ Note that τ_{gp1} results also from assuming filled deep traps and negligible Δn, or $\Delta p \sim \hat{p}_0 + \Delta n$.

‖ For 33 microseconds, as given in Table I of Hornbeck and Haynes,[54] calculated capture cross section for recombination would be smaller in proportion.

level, seems necessary. Otherwise, a near-saturation lifetime much larger even than τ_{gp1} would obtain.† From (60), (65) and (66) written for the higher level, the recombination lifetime will be $\tau_0 \sim \tau_{n0}$ throughout the small-signal range and with no small-signal saturation if $p_0 \gg C_{n3}n_3/C_{p3} = p_3^*$ and $p_0 \gg C_{n3}\mathfrak{N}_1/C_{p3} = (C_{p3}\tau_{n0})^{-1}$ hold. The latter condition‡ is $A_{p3} \gg 1.2 \times 10^{-17}$ cm^2. The coefficient C_{n3} and cross section A_{n3} for electron capture in the higher level are found to equal 2.5×10^{-8} cm^3 second^{-1} and 2.3×10^{-15} cm^2.

For the further analysis on the hypothesis that this model applies, the energy levels are properly treated as unknowns. The contribution $\tau_{gp1}(\tau_{tn1}/\tau_{gn1})$ to the time constant for decay in the shallower traps equals $\tau_\infty(\tau_{tn2}/\tau_{gn2}) = \tau_\infty(n_0 + n_2)/\hat{p}_0$ from the expression for τ_∞, and will accordingly be small compared with 10^{-2} second for deep traps sufficiently deep so that $n_0 + n_2 \ll 4 \times 10^8$ cm^{-3} holds.§ Then τ_{gp1} is 10^{-2} second and, with $p_1 \ll p_0$, C_{p1} has the value already found. An additional datum is available from experiment, namely the decay constant for the straggle effect: With the shallower traps nearly filled, multiple trapping results in an extended tail in the distribution of carriers from an injected pulse that are caused to drift past a fixed detector, at which the decay with time is measured. As shown in Section 3.4.3, the decay constant is the "straggle constant" ν_v, which is substantially $\nu_{gn1} + \nu_{tp1} + \nu_{gp1}$ for $p_0 \gg n_0$. Since ν_{gp1} is 10^2 second^{-1}, the observed value, 2×10^4 second^{-1}, is to be equated to $\nu_{gn1} \sim C_{n1}n_1$. With this result, the value for τ_{gp1} and $\mathfrak{N}_1 - \hat{n}_0 \sim \mathfrak{N}_1 = 2 \times 10^{12}$ cm^{-3}, the equation for τ_∞ contains only C_{n1} or n_1 and n_2 as unknowns. It fixes, say, $n_1/(n_0 + n_2)$ and thus approximately the separation between the energy levels, but there are not sufficient data with the model assumed to determine each level separately. It appears, however, from measurements relating to deep traps in samples of various conductivities,‖ that the location considered for these traps is substantially correct. With the trial value 2.8×10^6 cm^{-3} of n_2, the value obtained for C_{n1} is 1.2×10^{-6} cm^3 second^{-1}, and the value for n_1 is 1.7×10^{10} cm^{-3}, corresponding to an energy level for the shallower traps 0.007 ev above the trial location at the Fermi level in intrinsic material. With $\nu_{gn2} = 1$ second^{-1}, the value obtained for C_{n2} is 3.1×10^{-7} cm^3 second^{-1}. The cross sections $A_{n1} = 1.1 \times 10^{-13}$ cm^2 and $A_{n2} =$

† With appropriate notational changes, the result in the footnote on page 577 for (recombinative) deep traps applies to the shallower traps.

‡ This controls if the higher energy level is further than about 0.42 ev from the conduction band.

§ This condition holds by a factor of about 10^2 for the trial value of n_2.

‖ See Fig. 13 of Ref. 54.

2.9×10^{-14} cm^2 that result are half an order of magnitude smaller than the ones calculated by Hornbeck and Haynes.†

It seems likely that the photoconductivity under illumination intense enough to give the shorter decay times was quite appreciably offset as a result of heating of the sample.‡ Significant error from this source seems unlikely, however: The time constant for cooling was very probably comparable with the longer decay times, and measurements concerning these were made with considerably less intense initial illumination.§ There were presumably no pronounced effects of nonuniform generation in the thickness of the sample, the generation rate at the dark surface being at least 40 per cent of that at the illuminated surface, as calculated from the diffusion length for the shortest decay time.‖

Work has been done towards the identification of the impurities in silicon that occasion these trapping effects.[76,77] It might be noted that the energy levels suggest gold.[56,57] But there is evidence that gold gives a single center with two (or possibly more) levels, and such a center cannot account for the saturation of the shallower traps at a concentration less than that of the deep traps. Consider the assumption that a two-level model does apply, with shallow traps only partly filled in the experiments. Then τ_{gp1} applies for the decay at the shallower level and $\tau_{tn1}' \equiv 1/C_{n1}(\mathfrak{N} - \hat{n}_0)$ for the time constant of 20 microseconds observed with the spark source.¶ It follows that C_{n1} is 4×10^{-9} cm^3 second^{-1} for the $\mathfrak{N}$ of 1.15 $\hat{p}_0$ and the negligible $\hat{n}_0$ that the trial levels give. But, with this C_{n1}, the initial Δn immediately after the steady illumination that is shut

† Compared with the value from ten samples that they calculated in connection with Table I of their paper, A_{n2} is one order of magnitude smaller.

‡ Perhaps this heating accounts for apparent concentrations of normally empty traps determined from Fig. 4 of Ref. 54, which are about 0.7 of the values quoted.

§ Buck[74] has found a positive temperature coefficient of resistance in 38 ohm-cm p-type and 350 ohm-cm n-type silicon of 0.8 per cent per °C at room temperature, and has observed time constants for the cooling of the samples, similarly supported by wire leads and of comparable size and geometry, of the order of 100 seconds. The thermal time constant equals the heat capacity divided by the thermal dissipation constant, or power input per unit temperature elevation. For the sample here considered, 0.2 cm square and 2 cm long, power input is 8.7×10^{-4} watt for 10^{16} photons per cm^2 second absorbed, since 1 microwatt corresponds to $(5.1 \times 10^{16})\lambda$ photons per second of wavelength λ, and effective λ for the tungsten illumination is about 9×10^{-5} cm. The dissipation constant for a temperature elevation of 1°C with this power input in conjunction with the heat capacity of the sample of 0.14 joule per °C gives a thermal time constant of about 160 seconds. Haynes[75] has estimated a temperature elevation of no more than a few degrees for the more intense illuminations employed; heating of 3°C would decrease conductivity by an amount comparable with the total photoconductivity of Fig. 4 of Ref. 54.

‖ This diffusion length is 0.17 cm. In measurements on n-type silicon, a silicon filter and a constant-temperature enclosure were used.

¶ Note that τ_{tn1}' and $\tau_{tn1} \equiv 1/C_{n1}(\mathfrak{N} - \hat{n}_0 - \hat{p}_0)$ are the times for electron capture at the shallower level respectively for filled and empty deep traps.

off by the shutter is two orders of magnitude smaller than the apparent saturation value 2×10^{12} cm^{-3} of $\Delta\hat{n}$, and thus appreciably smaller than the initial value 6×10^{11} cm^{-3} of Δn estimated from the initial conductivity change.† Also, the apparent saturation value of $\Delta\hat{n}$ is proportional to Δg, which is not the fairly well-defined saturation observed.‡ Moreover, the value of C_{n1} gives a decay constant for the straggle effect, evaluated as $\nu_{gn1} + \nu_{gp1}$, equal to 157 second^{-1} and the same two orders of magnitude smaller than that observed. Inconsistencies largely similar result from the assumption of two-level trapping with recombination entirely in other centers.

No evidence has, indeed, been found that the observed trapping effects result through metallic impurities or through lattice defects produced either by mechanical deformation or by bombardment with high-energy electrons.[4,76,77,78] Present indications are that the deep traps in silicon are associated with the presence of oxygen as an impurity;[4,67,78] but these traps, as well as the specific reactions instrumental in their formation, have not yet been physically identified. Concentrations of the traps and of certain donor centers due to oxygen[79,80] have been found to be correlated.§ Both traps[76] and donors[85,86,87] are much more numerous in crystals grown (from quartz crucibles) with rotation of the seed than in those grown without, may be considerably increased in concentration by comparatively prolonged heating at 450°C, and may be largely removed quite rapidly by heating at temperatures above 500°C.‖ Concentrations of the shallower traps do not exhibit this dependence. The correlation is qualitative in that donor concentration is the more dependent on heating at 450°C; appreciable trap concentration may occur in an untreated crystal grown with rotation, and may assume a value considerably smaller than the donor concentration after heating at the lower temperature.[79,80,88]¶ It should further be noted that, while these observations have been mostly confined to n-type silicon (because the donors tend to convert p-type to n-type), observations concerning the deep and shallow traps which occur in p-type silicon indicate that a common mechanism is operative.[76]

† See Figs. 4 and 5 of Ref. 54.

‡ The steady-state equations give initial concentrations $\Delta n = \tau_{tn1}'(1 + \tau_{gp1}/\tau_{gn1})\Delta g$ and $\Delta\hat{n} = \tau_{gp1}\Delta g$. The apparent saturation value of $\Delta\hat{n}$ is the sum of these concentrations, since Δn is trapped rapidly, with time constant τ_{tn1}'.

§ Determinations of oxygen content from infrared absorption at 9 microns in combination with resistivity measurements on crystals heat treated at 450°C and 1000°C have shown that formation of these donors is associated with oxygen.[80-84]

‖ There seems to be an indication that the trap concentration is increased by water vapor but not by oxygen in the gaseous ambient.

¶ Deep traps originally present have been largely removed by heating only 5 seconds at 700°C and subsequently have been introduced in a concentration larger

3.3 *The Photomagnetoelectric Effect*

The steady-state effect with trapping will be analyzed on the basis of the general theory of Sections 2.1.1 and 2.1.2. For the PME field along an infinite slab to the faces of which the applied magnetic field is parallel and the y axis perpendicular, (28) and (32) give

$$E_x = \sigma^{-1}(I_x - \theta I_{Dy}) = \sigma^{-1}(I_x + \theta e D' d\Delta m/dy). \tag{111}$$

The total short-circuit current per unit width of slab along the magnetic field is accordingly given by

$$I_{sc} = \int_{-y_0}^{y_0} I_x \, dy = -\theta e \int_{-y_0}^{y_0} D' \frac{d\Delta m}{dy} \, dy, \tag{112}$$

and the field along the slab under the open-circuit condition is related to I_{sc} as previously derived.† To evaluate the integral in (112), Δm is first found from the continuity equation

$$d(D' d\Delta m/dy)/dy - \Delta m/\tau_m = 0, \tag{113}$$

which follows from (30) and (36); the drift term is either zero or of order θ^2 for the short-circuit or open-circuit condition. Since, for the slab, $I_{py} = -I_{ny} = I_{Dy}$, boundary conditions are

$$\begin{aligned} \mathfrak{L} - D' d\Delta m/dy &= s_{n1}\Delta n = s_{p1}\Delta p = s_{m1}\Delta m, \qquad y = y_0, \\ D' d\Delta m/dy &= s_{n2}\Delta n = s_{p2}\Delta p = s_{m2}\Delta m, \qquad y = -y_0, \end{aligned} \tag{114}$$

in which $\mathfrak{L}$ is the surface rate of generation of electron-hole pairs by strongly absorbed radiation and the right-hand members give surface recombination rates. For the linear small-signal case, the velocity functions s_n and s_p (with second subscripts "1" and "2" for the respective surfaces) are constants, with

$$s_m = (1 - r_n)s_n = (1 - r_p)s_p \tag{115}$$

the surface recombination velocity for Δm.

The increase in conductance of the slab is given by

$$\Delta G = e \int_{-y_0}^{y_0} (\mu_n \Delta n + \mu_p \Delta p) \, dy = e(\mu_n + \mu_p) \int_{-y_0}^{y_0} (\tau_c/\tau_m)\Delta m \, dy. \tag{116}$$

The second form follows from $\Delta n/\tau_n = \Delta p/\tau_p = \Delta m/\tau_m = \mathfrak{R}_m$, with

than the original one by heating 16 hours at 470°C.[76] It is not yet known whether prolonged heating at 1000°C, which prevents appreciable subsequent introduction of donors at the lower temperature[82,85,86,87] would similarly prevent the introduction of deep traps.

† See Ref. 11, Equation (39).

$$\tau_c \equiv \Delta\sigma/e(\mu_n + \mu_p)\Delta g = (\mu_n\tau_n + \mu_p\tau_p)/(\mu_n + \mu_p) \quad (117)$$

a lifetime function that determines the conductivity increase $\Delta\sigma$ for the uniform volume generation rate $\Delta g = \mathfrak{R}_m$. For the linear small-signal case, the lifetime functions are constants and

$$\begin{aligned} \Delta G &= e(\mu_n + \mu_p)\tau_c D_0'(d\Delta m/dy)\,|_{y=-y_0}^{y=y_0} \\ &= e(\mu_n + \mu_p)\tau_c(\mathfrak{L} - s_{m1}\Delta m\,|_{y=y_0} - s_{m2}\Delta m\,|_{y=-y_0}) \end{aligned} \quad (118)$$

follows by use of (113) and (114).

These results show how the theory previously given for the PME effect without trapping[11] is readily generalized to include trapping by writing equations in terms of Δm and the diffusivity D', lifetime function τ_m and surface recombination function s_m, and employing suitably generalized ΔG. Experiment may determine D_0', τ_m and s_m. In accordance with (115) and the results of Section 2.1.2, each of these gives rise, as determined by the trapping ratios, to corresponding quantities for electrons and for holes.

In its dependence on trapping and recombination in centers of a single type, the PME effect is generally nonlinear if deep traps in the minority-carrier trapping range are involved. Then trap saturation occurs in the small-signal range, as described in Section 3.1.2, and the lifetime may be nonuniform: From the illuminated surface into the slab, it may decrease from a saturation value to a much smaller linear small-signal value, a transition value at a given depth being sharply dependent on light intensity.

The influence on the PME effect of trapping as such may be investigated by assuming traps that may be nonrecombinative in conjunction with recombination on the dark surface, or with recombination in the volume of the idealized type discussed in Section 3.2.2. With the latter procedure, the linear recombination term $(-\nu_{n3}\Delta n - \nu_{p3}\Delta p)$ is included in the continuity equation. For the linear small-signal case, τ_m for the traps is thus replaced by $\bar{\tau}_m \equiv [\tau_m^{-1} + (1 - r_n)\nu_{n3} + (1 - r_p)\nu_{p3}]^{-1}$. For p-type material, ν_{p3} is set equal to zero, and $\tau_3 \equiv \nu_{n3}^{-1}$ is introduced. Then, for nonrecombinative electron traps of the acceptor type, $r_n = \nu_{tn1}/(\nu_{tn1} + \nu_{gn1})$, $r_p = 0$, $D_0' = [1 - p_0\mathfrak{N}_1^*/(n_0 + p_0)(\mathfrak{N}_1^* + n_0)]D_0$ and $\bar{\tau}_m = (1 + \nu_{tn1}/\nu_{gn1})\tau_3$ are obtained by use of the first of (31) and (60). If the traps are of the donor type, then $r_n = 0$, $r_p = -\nu_{tn2}/\nu_{gn2}$, $D_0' = [1 + \mathfrak{N}_2^*/(n_0 + p_0)]D_0$ and $\bar{\tau}_m = \tau_3$ are obtained. Essentially the same diffusion-length lifetime associated with D_0, namely

$$\bar{\tau}_0 = [1 + \mathfrak{N}_j^*/(n_0 + p_0)]\tau_3 \equiv K_\tau\tau_3, \qquad j = 1, 2, \quad (119)$$

results for both types of traps. Thus, minority-carrier trapping increases the diffusion-length lifetime and hence decreases PME current. The effect is appreciable in cases for which the capture concentration $\mathfrak{N}_j^*$ is at least comparable with the equilibrium concentration of majority carriers.† Similar analysis for nonrecombinative majority-carrier traps gives a K_τ which is that of (119) modified by division by $1 + \mathfrak{N}_j^*/n_0$ for n-type material or by $1 + \mathfrak{N}_j^*/p_0$ for p-type. Thus, majority-carrier trapping decreases the diffusion-length lifetime and increases PME current, but this increase is only that for a K_τ no smaller than $(1 + p_0/n_0)^{-1}$ or $(1 + n_0/p_0)^{-1}$, respectively.

Capture cross sections, concentrations and energy levels of traps may be found from suitable PME and photoconductivity measurements at a single temperature. Theory for trapping and recombination in traps of a single type, which holds whatever the method be for determining diffusion length L_0 and lifetime τ_0 will be considered first; while the PME method has certain advantages, any one of a number of other methods may also be employed.‡ In view of the fundamental restriction of (53) to which the four capture and release frequencies are subject, it will be convenient to deal with the capture frequencies ν_{tn1} and ν_{tp1} and the capture concentration $\mathfrak{N}_1^*$ of (63) as independent parameters. To determine these parameters, three quantities must be measured. Suppose, for example, that from suitable linear small-signal measurements, τ_0, τ_c and the lifetime τ_2 of (92) for decay of photoconductivity are known. Solving (60) and (92) for τ_{tn1}, τ_{tp1} and $\mathfrak{N}_1^*$ gives§

$$\begin{aligned} \tau_{tn1} &= [(\tau_2 - \tau_p)/(\tau_2 - \tau_a)]\tau_0, \\ \tau_{tp1} &= [(\tau_2 - \tau_n)/(\tau_2 - \tau_a)]\tau_0, \\ \mathfrak{N}_1^* &= (n_0 + p_0)(\tau_2 - \tau_a)/(\tau_n + \tau_p - \tau_2), \end{aligned} \tag{120}$$

in which τ_a is defined by

$$\tau_a \equiv (n_0\tau_n + p_0\tau_p)/(n_0 + p_0). \tag{121}$$

Then, with

$$\begin{aligned} \tau_n &= [(\mu_n + \mu_p)n_0\tau_c - \mu_p(n_0 + p_0)\tau_0]/(\mu_n n_0 - \mu_p p_0), \\ \tau_p &= [\mu_n(n_0 + p_0)\tau_0 - (\mu_n + \mu_p)p_0\tau_c]/(\mu_n n_0 - \mu_p p_0), \end{aligned} \tag{122}$$

† It can be shown that, if different types of traps are present, the $\mathfrak{N}_j^*$ in (119) is replaced by the sum of the respective capture concentrations.

‡ See, for example, van Roosbroeck and Buck.[89]

§ Note, from (60) and (92), that τ_2 is larger than τ_n, τ_p and τ_a, and smaller than $\tau_n + \tau_p$.

subject to† $\mu_n n_0 \neq \mu_p p_0$, which are obtained by solving the equations defining τ_0 and τ_c, (35) and (117), τ_{tn1}, τ_{tp1} and $\mathfrak{N}_1^*$ can be found from experiment. It should be noted that (117) does not represent an advantageous method for determining τ_c; an indirect method will be given that obviates the necessity of knowing the intensity of absorbed radiation.

An additional independent datum is required to determine C_{n1}, C_{p1}, $\mathfrak{N}_1$ and n_1 or p_1. As (41), (42), (50) and (63) show, τ_{n0}, τ_{p0}, or equilibrium concentration of empty traps would serve.‡ Thus, a measurement involving the saturation range is required in addition to those in the linear small-signal range. It is, in fact, desirable that two such measurements be made, for reasons that will be discussed. Suppose, for example, that there is small-signal trap saturation in p-type material and that the decay frequency $C_{p1}\mathfrak{N}_1 = \tau_{p0}^{-1}$ in the saturation range and $\mathfrak{N}_1 - \hat{n}_0$ are known in addition to τ_0 and τ_2. It follows then, from the first equation of (63), that $\mathfrak{N}_1^*$ is ν_{tp1} times a known constant:

$$\begin{aligned} \mathfrak{N}_1^* &= (C'_{p1}\tau_{tp1})^{-1}, \\ C'_{p1} &\equiv \tau_{p0}^{-1}(\mathfrak{N}_1 - \hat{n}_0)^{-1} = (1 + p_1/p_0)C_{p1}. \end{aligned} \tag{123}$$

Eliminating $n_0\tau_{tp1} + p_0\tau_{tn1}$ from τ_2 of (92) by use of the third form for τ_0 of (65) results in an equation linear in τ_{tn1} and τ_{tp1}, after $\mathfrak{N}_1^*$ has been eliminated by use of (123). This linear equation and the one for τ_0 may be solved for τ_{tn1} and τ_{tp1}, and, with (50), this solution gives

$$\begin{aligned} \hat{n}_0/\mathfrak{N}_1 &= (1 + n_1/n_0)^{-1} \\ &= \tau_{p0}\frac{(p_0 - n_0)/(p_0 + n_0) - C'_{p1}p_0(\tau_2 - \tau_0)}{p_0\tau_2/(p_0 + n_0) - \tau_0}, \\ C_{n1} &= (\mathfrak{N}_1 - \hat{n}_0)^{-1} \\ &\quad \cdot \frac{(p_0 - n_0)/(p_0 + n_0) - C'_{p1}p_0(\tau_2 - \tau_0)}{\tau_0 - n_0\tau_2/(p_0 + n_0) - C'_{p1}(p_0 + n_0)(\tau_2 - \tau_0)\tau_0}, \\ C_{p1} &= C'_{p1} - \tau_{p0}^{-1}(\mathfrak{N}_1 - \hat{n}_0)^{-1}\hat{n}_0/\mathfrak{N}_1. \end{aligned} \tag{124}$$

With trapping, the PME current-conductance ratio does not determine τ_0 but depends also on τ_c (which differs from τ_0 because of trapping), and direct determination of τ_c requires knowledge of light intensity.

† With the denominator $\mu_n n_0 - \mu_p p_0$ equal to zero, $\tau_0 = \tau_c$ follows, and the numerators are also zero.

‡ The neutrality condition would serve in cases of trapping by acceptors which determine the (p-type) conductivity, for which $\hat{n}_0 = p_0 - n_0$ holds; but $\hat{n}_0$ and $p_0 - n_0$ that differ phenomenologically must generally be considered to obtain.

This circumstance need not, however, really vitiate what in the no-trapping case is a primary motivation for dealing with this ratio: What is determined independently of light intensity is a relationship between τ_0 and τ_c, and a further relationship between these lifetimes will serve to determine both. For C'_{p1} known, the relationship

$$C'_{p1} = (\tau_{tp1}\mathfrak{N}_1^*)^{-1} = (\tau_n + \tau_p - \tau_2)/(p_0 + n_0)(\tau_2 - \tau_n)\tau_0 \quad (125)$$

obtained from (120) is, with (122), linear in τ_c and readily solved for this lifetime.

The PME method of the high-recombination-velocity dark surface is best employed, since it generally provides better accuracy for the conductance change than does the thick-slab method which it otherwise subsumes as a limiting case.† Optimum slab thickness is about one or two diffusion lengths. For large dark-surface recombination velocity, the small-signal results for no trapping‡ give, for the present case,

$$\begin{aligned} I_{sc} &= -\theta e \mathfrak{L} L_0 (S_1 + \coth 2Y_0)^{-1} \\ &= -\theta(\mu_n + \mu_p)^{-1}(L_0/\tau_c)(\coth Y_0)\Delta G, \end{aligned} \quad (126)$$

in which diffusion length L_0 is $(D_0\tau_0)^{\frac{1}{2}}$, Y_0 is y_0/L_0 and S_1 is $s_{m1}L_0/D_0'$; note that ΔG now involves τ_c as a factor. Thus, τ_c is given by

$$\tau_c = \frac{2Y_0 \coth Y_0}{[\mathfrak{I}/(\Delta G/G_0)]}\tau_0, \quad (127)$$

where

$$\mathfrak{I}/(\Delta G/G_0) \equiv -2y_0(\mu_n + \mu_p)I_{sc}/\theta D_0 \Delta G \quad (128)$$

is the dimensionless PME current-conductance ratio. In (127), τ_0 enters also through Y_0, and, with (125), both τ_0 and τ_c may be found. Note that apparent lifetime τ_r on the assumption of no trapping, obtained by equating $\mathfrak{I}/(\Delta G/G_0)$ to $[2y_0/(D_0\tau_r)^{\frac{1}{2}}] \coth [y_0/(D_0\tau_r)^{\frac{1}{2}}]$, is related to τ_0 and τ_c by $\tau_r \tanh^2 [y_0/(D_0\tau_r)^{\frac{1}{2}}] = (\tau_c^2/\tau_0) \tanh^2 Y_0$, and equals τ_c^2/τ_0 only for the thick slab, for which the hyperbolic tangents are unity.

If the model that applies is that of nonrecombinative traps with recombination in other centers, then (119) gives the lifetime $\bar{\tau}_0$ upon which the linear small-signal I_{sc} for minority-carrier trapping depends. For ΔG, (116) holds for the linear small-signal case, for which τ_c/τ_m is $[(1 - r_n)\mu_n + (1 - r_p)\mu_p]/(\mu_n + \mu_p)$. The solution for Δm is readily obtained by comparison with that for the corresponding no-trapping

† See Ref. 11, Section 3.42.
‡ See Ref. 11, Equation (50).

case.† With the factor L_0^2/D_0' replaced by $\bar{\tau}_m$, it is found that

$$\Delta G = \frac{K_G e(\mu_n + \mu_p)\tau_3 \mathcal{E}(\cosh 2Y_0 - 1)}{S_1 \sinh 2Y_0 + \cosh 2Y_0} \tag{129}$$

is the linear small-signal ΔG for large dark-surface recombination velocity, in which Y_0 and S_1 are defined as above, but in terms of $L_0 \equiv (D_0\bar{\tau}_0)^{\frac{1}{2}}$. The factor $K_G \equiv (\tau_c/\tau_m)(\bar{\tau}_m/\tau_3)$ is τ_c/τ_n in general for p-type material, for which $\bar{\tau}_m$ is $\tau_3/(1 - r_n)$; for n-type material, K_G is τ_c/τ_p. The expression that K_G multiplies also depends on trapping, since L_0 does. Equations (60) and (117) give, for p-type material,‡

$$\begin{aligned} K_G &= 1 + (b + 1)^{-1}\mathfrak{N}_j^*/n_0 \qquad \text{(electron trapping)} \\ K_G &= \frac{1 + (1 + b^{-1})^{-1}\mathfrak{N}_j^*/p_0}{1 + \mathfrak{N}_j^*/p_0} \qquad \text{(hole trapping).} \end{aligned} \tag{130}$$

For hole and electron trapping, respectively, in n-type material, n_0 and p_0 in these equations are interchanged and $b \equiv \mu_n/\mu_p$ replaced by its reciprocal.

With I_{sc} for this model given by the first form of (126) with the redefined L_0,

$$\mathcal{I}/(\Delta G/G_0) = (K_\tau/K_G)\, 2Y_0 \coth Y_0 \tag{131}$$

follows by use of (119), (128) and (129). Apparent lifetime τ_r is accordingly given by $\tau_r \tanh^2 y_0/(D_0\tau_r)^{\frac{1}{2}} = (K_G^2/K_\tau)\tau_3 \tanh^2 Y_0$, and equals $(K_G^2/K_\tau)\tau_3$ for the thick slab. As trap concentration increases, diffusion length increases and a slab of any given thickness becomes a "thin" slab, for which $Y_0 \coth Y_0 \sim 1$; and $\mathcal{I}/(\Delta G/G_0)$ approaches a constant value that is independent of the thickness. For example, if the half-thickness y_0 is of order $(D_0\tau_3)^{\frac{1}{2}}$, then $K_\tau \gg 1$ or $\mathfrak{N}_j^* \gg n_0 + p_0$ also gives small Y_0. From the expressions for K_τ and (130) for K_G it is found that $\mathcal{I}(\Delta G/G_0)$ approaches $2(b + 1)n_0/(n_0 + p_0)$ for electron trapping and $2(b + 1)p_0/b(n_0 + p_0)$ for hole trapping, regardless of conductivity type. On the other hand, if the slab is so thick that $y_0^2 \gg D_0\tau_3\mathfrak{N}_j^*/(n_0 + p_0)$ holds, then the condition $\mathfrak{N}_j^* \gg n_0 + p_0$ for large

† In Equation (44) of Ref. 11, Δp is replaced by Δm; the D_0 that appears explicitly originates from the boundary conditions and is replaced by D_0'; and S_1 and S_2 are the velocities for Δm multiplied by L_0/D_0'.

‡ Note that K_τ and K_G are equal (for electron or hole trapping) in p-type material for which $\mu_n n_0 = \mu_p p_0$ holds.

trap concentration gives† τ_3/τ_r equal to K_r/K_G^2 or $(b+1)^2n_0^2/(n_0+p_0)\mathfrak{N}_j^*$ for electron trapping and $(b+1)^2p_0/b^2(n_0+p_0)$ for hole trapping in p-type material, with similar results for n-type obtained as in connection with (130).

With the aid of suitable saturation-range measurements, the cross section, concentration and energy level of the nonrecombinative traps are readily found. For traps of the acceptor type, C_{n1}, $\mathfrak{N}_1$ and n_1 or p_1 are to be determined, and these can easily be calculated from values from experiment of $\mathfrak{N}_1 - \hat{n}_0$, saturation-range lifetime τ_3, lifetime $\tau_{gn1} \equiv [C_{n1}(n_0+n_1)]^{-1}$ for the traps nearly full, and any one of τ_0, τ_∞ or $\mathfrak{N}_1^*$. These last three quantities are not independent; from (107), lifetime τ_∞ for the traps nearly empty is $[1+(\mathfrak{N}_1-\hat{n}_0)/(n_0+n_1)]\tau_3 = [1+\mathfrak{N}_1^*/n_0]\tau_3$. Measurement of $g/(\Delta G/G_0)$ serves to determine $\bar{\tau}_0$: By means of (119) and (130), K_r/K_G may be written as $(b+1)n_0[n_0+p_0-(p_0-bn_0)\tau_3/\bar{\tau}_0]^{-1}$ for electron trapping in p-type material, or as an analogous expression for hole trapping in n-type, so that (131) involves only $\bar{\tau}_0$ as unknown.

3.4 *Transport of Injected Carriers*

3.4.1 *The Linear Differential Equations*

The general differential equations of Sections 2.1 and 2.2 are here specialized to the linear small-signal case of trapping (and recombination) in centers of a single type, for which certain specific transport problems will be considered. From (6) through (9) and (46), the linear continuity equation for centers of the acceptor type is

$$\begin{aligned}\partial\Delta m/\partial t &= \hat{D}_p \operatorname{div}\operatorname{grad}\Delta p + \hat{D}_n \operatorname{div}\operatorname{grad}\Delta n \\ &\quad -\hat{\mathbf{v}}_p\cdot\operatorname{grad}\Delta p - \hat{\mathbf{v}}_n\cdot\operatorname{grad}\Delta n + \Delta g - \mathfrak{R}_m \\ &= \partial\Delta p/\partial t = D_0 \operatorname{div}\operatorname{grad}\Delta p - \hat{D}_n \operatorname{div}\operatorname{grad}\Delta\hat{n} \\ &\quad - \mathbf{v}_0\cdot\operatorname{grad}\Delta p + \hat{\mathbf{v}}_n\cdot\operatorname{grad}\Delta\hat{n} + \Delta g + \nu_{11}\Delta p + \nu_{12}\Delta\hat{n},\end{aligned} \tag{132}$$

the first form being that which applies for the linear case in general. The diffusivity and velocity with minority-carrier subscripts of those defined by

† It has been shown by Amith[62,63,90] that, for minority-carrier trapping in the thick slab, τ_3/τ_r is proportional to $\mathfrak{N}_j^{-2}$ for large $\mathfrak{N}_j$, if K_r is taken as unity. This dependence obtains in the intermediate range in which $\mathfrak{N}_j^*$ is large compared with minority-carrier concentration n_0 or p_0 but small compared with p_0 or n_0 so that the change in diffusion length may be neglected. For majority-carrier trapping in general, τ_3 and τ_r are substantially equal in this range.

$$\begin{aligned} \hat{D}_p &\equiv n_0 D_0/(n_0 + p_0), \qquad \hat{D}_n \equiv p_0 D_0/(n_0 + p_0), \\ \hat{\mathbf{v}}_p &\equiv n_0 \mathbf{v}_0/(n_0 - p_0), \qquad \hat{\mathbf{v}}_n \equiv -p_0 \mathbf{v}_0/(n_0 - p_0) \end{aligned} \tag{133}$$

are, in sufficiently strongly extrinsic material, substantially the minority-carrier diffusivity and velocity, as are D_0 and $\mathbf{v}_0$; while those with majority-carrier subscripts are comparatively small. The linear equation

$$\partial \Delta \hat{n}/\partial t = \nu_{21} \Delta p + \nu_{22} \Delta \hat{n} \tag{134}$$

holds for interactions with the traps. Eliminating $\Delta \hat{n}$ from (132) and (134), substituting from (133) and making use of (51), (62) and (63) results in

$$\begin{aligned} \partial^2 \Delta p/\partial t^2 - D_0 \operatorname{div} \operatorname{grad}(\partial \Delta p/\partial t) + \mathbf{v}_0 \cdot \operatorname{grad}(\partial \Delta p/\partial t) + \nu_s \partial \Delta p/\partial t \\ - \nu_D D_0 \operatorname{div} \operatorname{grad} \Delta p + \nu_v \mathbf{v}_0 \cdot \operatorname{grad} \Delta p + \Delta_1 \Delta p \\ = \partial \Delta g/\partial t + (\nu_s - \nu_{tp1}) \Delta g, \end{aligned} \tag{135}$$

with ν_s defined by (83), and ν_D and ν_v by

$$\begin{aligned} \nu_D &\equiv [1 + \mathfrak{N}_1^*/(n_0 + p_0)](\nu_{gn1} + \nu_{gp1}) \\ \nu_v &\equiv \nu_{gn1} + \nu_{gp1} + (\nu_{gn1} - \nu_{gp1}) \mathfrak{N}_1^*/(n_0 - p_0). \end{aligned} \tag{136}$$

The frequency ν_v will be referred to as the "straggle constant". It is readily shown that the linear differential equations that Δn and $\Delta \hat{n}$ satisfy are entirely similar to (135) except for suitable modifications of the right-hand member; all the concentrations satisfy the same equation if there is no volume generation. For Δn it suffices to replace ν_{tp1} where it occurs explicitly by ν_{tn1} , while for $\Delta \hat{n}$ only the generation term $(\nu_{tn1} - \nu_{tp1}) \Delta g$ occurs, $\partial \Delta g/\partial t$ being absent. It is also readily shown that linear recombination in other centers can be taken into account by adding $\nu_{n3} + \nu_{p3}$ to the coefficient ν_s of $\partial \Delta n/\partial t$ and $\partial \Delta p/\partial t$ and $(\nu_s - \nu_{tn1}) \nu_{n3} + (\nu_s - \nu_{tp1}) \nu_{p3}$ to the coefficient Δ_1 of Δn and Δp.

3.4.2 *Steady-State Transport; Reverse Drift*

A simple case that yields qualitative information of interest is that of injection into a filament in the steady state with applied field. For this case,

$$\nu_D D_0 d^2 \Delta p/dx^2 - \nu_v v_0 d\Delta p/dx - \Delta_1 \Delta p = 0 \tag{137}$$

is to be solved for, say, Δp zero for distance x along the filament negatively or positively infinite and continuous at the origin at which there

is carrier injection with zero injected total-current density. Equation (137) is easily shown to be equivalent to (30) and (31) specialized for no volume generation and acceptor centers only; ν_D and ν_v are $(-\nu_{22}) = \nu_s - \nu_{tp1}$ times D_0'/D_0 and v_0'/v_0, respectively, and τ_p, from (60), is $(\nu_s - \nu_{tp1})/\Delta_1$.

The solutions in the semi-infinite regions separated by the origin are $e^{r_1 x}$ and $e^{r_2 x}$, where r_1 and r_2 are given by

$$\begin{aligned}\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} &= \tfrac{1}{2}\left[\nu_v v_0/\nu_D D_0 \pm \left[(\nu_v v_0/\nu_D D_0)^2 + 4\Delta_1/\nu_D D_0\right]^{\frac{1}{2}}\right] \\ &\sim \begin{pmatrix} \nu_v v_0/\nu_D D_0 \\ -\Delta_1/\nu_v v_0 \end{pmatrix},\end{aligned} \tag{138}$$

as obtained from (137). The case of recombination without appreciable trapping[64] presents no unfamiliar features; the approximation given, which is that for Δ_1 small, as may result from one of the capture coefficients small, will accordingly be considered. The magnitude of r_1 is thus large compared with that of r_2. With the condition $v_0 > 0$, which may be assumed without loss of generality, $e^{r_1 x}$ gives the familiar sharply varying field-opposing solution to the left of the origin and $e^{r_2 x}$ gives the corresponding gradually varying field-aiding solution to the right, provided ν_v is positive; then, r_1 and r_2 are respectively positive and negative. But negative ν_v can occur, for which an anomalous behavior obtains, the field-opposing and field-aiding solutions then being respectively the gradually and sharply varying exponentials $e^{r_2 x}$ and $e^{r_1 x}$. For this case, in the limit of no diffusion, added carrier concentration appears only in the direction opposite to that of the ambipolar drift velocity, that is, opposite to the direction of drift normally determined by conductivity type.

This "reverse drift" associated with trapping may be understood in terms of properties of the current density $\Delta\mathbf{I}$ of added carriers. From (19), added carriers drift in the direction of the total current density, or the contribution to $\Delta\mathbf{I}$ from drift has the sign of $\mathbf{I}$ if $n_0\Delta p - p_0\Delta n$ or $\Delta p/\Delta n - p_0/n_0$ is positive; that is, if injection results in proportionately more holes than electrons than is the case at thermal equilibrium. This behavior is, of course, that which normally occurs in n-type material; with no trapping, $\Delta p/\Delta n$ equals unity and added carriers drift with or opposite to $\mathbf{I}$ according to whether the semiconductor is n-type or p-type, with no drift in intrinsic material.[10,64] Thus, the normal behavior requires the conditions that $\Delta p/\Delta n - p_0/n_0$ be positive in n-type and negative in p-type. It is easily shown, by writing these conditions by means of (60) for the steady-state value $(1 - r_n)^{-1}$ of $\Delta p/\Delta n$, that

both are tantamount in the steady state to the single condition, $\nu_v > 0$. This condition clearly always holds for the majority-carrier trapping range, while reverse drift results for sufficient minority-carrier trapping in not too strongly extrinsic material. From (136), $\nu_v > 0$ gives

$$p_0 - n_0 > (\nu_{gn1} - \nu_{gp1})\mathfrak{N}_1^*/(\nu_{gn1} + \nu_{gp1}) \tag{139}$$

for p-type material, and a similar inequality for n-type is obtainable by changing the sign of each side. Equating the two sides gives the condition for no drift, which, for no trapping, holds for intrinsic material. From (54), the right-hand side of (139) may be written as

$$(n_0 - n_1^*)\mathfrak{N}_1^*/(n_0 + n_1^*) = (p_1^* - p_0)\mathfrak{N}_1^*/(p_1^* + p_0).$$

It reduces to $\mathfrak{N}_1^*$ for electron trapping without recombination, since then ν_{gp1} and n_1^* are zero. For this case, since $\mathfrak{N}_1^*$ equals $n_0\nu_{tn1}/\nu_{gn1}$ from (63), reverse drift obtains if n_0/p_0 in p-type material exceeds $\tau_{tn1}/(\tau_{tn1} + \tau_{gn1})$, the fraction of the time electrons are free. A similar result holds for hole trapping in n-type material.

3.4.3 *Drift of an Injected Pulse*

The differential equation for drift with negligible diffusion and no volume generation in one cartesian dimension with trapping by centers of a single type is

$$\partial^2\Delta p/\partial t^2 + v_0\partial^2\Delta p/\partial x\partial t + \nu_s\partial\Delta p/\partial t + \nu_v v_0\partial\Delta p/\partial x + \Delta_1\Delta p = 0, \tag{140}$$

from (135). For a pulse of carriers injected into a doubly infinite filament, a suitable technique of solution is that of the bilateral or two-sided Laplace transform[91,92] with respect to the distance variable, for which the notations

$$F(s, U) \equiv \int_{-\infty}^{\infty} e^{-s\lambda} f(\lambda, U)\, d\lambda \equiv \mathfrak{L}\{f(X, U)\} \equiv \overline{f(X, U)} \tag{141}$$

are here employed. Dimensionless independent variables

$$X \equiv x/L, \qquad U \equiv t/\tau \tag{142}$$

are introduced, and with distance and time units given by†

$$\begin{aligned} L &\equiv v_0\tau, \\ \tau &\equiv (\,|\,\nu^2\,|\,)^{-\frac{1}{2}}, \\ \nu^2 &\equiv 4[\nu_v(\nu_s - \nu_v) - \Delta_1] \\ &\equiv 4n_i^2(\nu_{tn1} - \nu_{tp1})^2(p_0 - n_0)^{-2} \\ &\quad \cdot[(\nu_{tn1} + \nu_{tp1})(p_0 - n_0)/(\nu_{tn1} - \nu_{tp1})\mathfrak{N}_1^* - 1], \end{aligned} \tag{143}$$

† The second form for ν^2 follows by use of (52), (62) and (63).

subject to the restrictions $\nu^2 \neq 0$ and $n_0 \neq p_0$, the reduced equation

$$\begin{aligned} &\partial^2\Delta p/\partial U^2 + \partial^2\Delta p/\partial X\partial U + \zeta\partial\Delta p/\partial U \\ &\qquad + \tfrac{1}{2}(\zeta + \kappa)\partial\Delta p/\partial X + \tfrac{1}{4}(\zeta^2 - \kappa^2 \mp 1)\Delta p = 0 \end{aligned} \tag{144}$$

results, where κ and ζ are the parameters

$$\begin{aligned} \kappa &\equiv (2\nu_v - \nu_s)\tau \\ &= [\nu_{gn1} + \nu_{gp1} + (n_0 + p_0)(\nu_{tn1} - \nu_{tp1})/(n_0 - p_0)]\tau, \\ \zeta &\equiv \nu_s\tau = (\nu_{tn1} + \nu_{gn1} + \nu_{tp1} + \nu_{gp1})\tau > 0. \end{aligned} \tag{145}$$

Coefficient unity for the second term of (144) results from the definition of L. The double sign in the last term of the equation results from the necessity of defining a real (and positive) τ, the upper and lower signs applying respectively for positive and negative ν^2.

Laplace transformation of (144) gives

$$\begin{aligned} &d^2\overline{\Delta p}/dU^2 + (s + \zeta)d\overline{\Delta p}/dU \\ &\qquad + [\tfrac{1}{2}(\zeta + \kappa)s + \tfrac{1}{4}(\zeta^2 - \kappa^2 \mp 1)]\overline{\Delta p} = 0. \end{aligned} \tag{146}$$

As has governed the choices of L and τ, the roots $(-N_1)$ and $(-N_2)$ of the associated quadratic reduce to

$$\begin{pmatrix} -N_1 \\ -N_2 \end{pmatrix} = \tfrac{1}{2}\{\pm[(s - \kappa)^2 \pm 1]^{\frac{1}{2}} - (s + \zeta)\}, \tag{147}$$

in which the double sign inside the radical here and in what follows relates only to the sign of ν^2. Equation (146) holds for each of the transformed concentrations, as does (140) for each of the original ones. General solutions are thus

$$\begin{aligned} \overline{\Delta p} &= \sum_{j=1}^{2} A_j e^{-N_j U}, \\ \overline{\Delta \hat{n}} &= \sum_{j=1}^{2} r_{nj} A_j e^{-N_j U}. \end{aligned} \tag{148}$$

From these solutions in conjunction with the Laplace transform of $\partial\Delta\hat{n}/\partial U$ obtained from (134), it is found that the r_{nj} are given by

$$\begin{aligned} r_{nj} &= -\nu_{21}/(\nu_{22} + N_j/\tau) \\ &= (\nu_{tn1} - \nu_{tp1})/(\nu_{tn1} + \nu_{gn1} + \nu_{gp1} - N_j/\tau), \qquad j = 1,2. \end{aligned} \tag{149}$$

With ν_j replaced by N_j/τ, these are formally the same as the trapping ratios for the decay of photoconductivity given in the first line of (81).†

† Other forms for the r_{nj} obtainable through Laplace transformation of $\partial\Delta p/\partial t$ from (132) written for one dimension and no diffusion are not similarly related to the forms of the second line of (81), though, for $s = 0$, the N_j/τ reduce to the ν_j and all r_{nj} to those for photoconductivity.

The A_j are easily found in terms of the r_{nj} and transforms $\overline{\Delta p_1}$ and $\overline{\Delta \hat{n}_1}$ of the initial distributions. For no carriers trapped initially, the case that will be considered, (147) to (149) give

$$\begin{pmatrix} A_1/\overline{\Delta p_1} \\ A_2/\overline{\Delta p_1} \end{pmatrix} = \tfrac{1}{2}\{1 \mp (s - \xi)/[(s - \kappa)^2 \pm 1]^{\frac{1}{2}}\}, \tag{150}$$

in which the parameter ξ is defined by

$$\xi \equiv (\nu_{11} - \nu_{22})\tau = (\nu_s - 2\nu_{tp1})\tau. \tag{151}$$

Corresponding coefficients for the solution for $\Delta\bar{n}$ are $(1 - r_{n1})A_1$ and $(1 - r_{n2})A_2$, and these are similar to the ones for $\Delta\bar{p}$ if ξ in (150) is replaced by

$$\eta \equiv (\nu_{11} - \nu_{22} - 2\nu_{21})\tau = (\nu_s - 2\nu_{tn1})\tau. \tag{152}$$

Solutions for an injected gaussian delta pulse are advantageously derived as limiting forms as a approaches zero of solutions for the gaussian initial distribution whose transform is given by

$$\overline{\Delta p_1}/(\mathcal{P}/L) = \mathfrak{L}\{\tfrac{1}{2}\pi^{-\frac{1}{2}}a^{-1}e^{-X^2/4a^2}\} = e^{a^2s^2} \tag{153}$$

for $\mathcal{P}$ carrier pairs injected per unit area of cross section. From (147) through (153), $\overline{\Delta p}$ for this initial gaussian distribution is given by

$$\begin{aligned}\overline{\Delta P} \equiv \overline{\Delta p}/(\mathcal{P}/L) = e^{\,a^2s^2 - \frac{1}{2}U(s+\xi)}\,\big(&\cosh\{\tfrac{1}{2}U[(s - \kappa)^2 \pm 1]^{\frac{1}{2}}\} - (s - \xi) \\ &\cdot[(s - \kappa)^2 \pm 1]^{-\frac{1}{2}} \sinh\{\tfrac{1}{2}U[(s - \kappa)^2 \pm 1]^{\frac{1}{2}}\}\big),\end{aligned} \tag{154}$$

from which $\overline{\Delta N} \equiv \overline{\Delta n}/(\mathcal{P}/L)$ is obtained by replacing ξ by η.

Certain inverse Laplace transforms that are needed are derived in Appendix A, and Appendix B includes some details of their use in calculation of the solutions from (154). Solutions for the initial delta pulse at the origin are found to be

$$\begin{aligned}\Delta P \equiv \Delta p/(\mathcal{P}/L) = {}& [e^{\,\kappa X - \frac{1}{2}(\xi+\kappa)U}] \\ &\cdot\Big(\delta(U - X) + \frac{1}{2}\Big\{(\xi - \kappa)\,\frac{I_0}{J_0}[\sqrt{X(U - X)}] \\ &+ I_1[\sqrt{X(U - X)}]\,\frac{X}{\sqrt{X(U - X)}}\Big\}\,\mathbf{1}\,[X(U - X]\Big),\end{aligned} \tag{155}$$

$$\begin{aligned}\Delta\hat{N} \equiv \Delta\hat{n}/(\mathcal{P}/L) = {}& \tfrac{1}{2}(\xi - \eta)[e^{\kappa X - \frac{1}{2}(\xi+\kappa)U}] \\ &\cdot\frac{I_0}{J_0}[\sqrt{X(U - X)}]\,\mathbf{1}\,[X(U - X)];\end{aligned}$$

for $\Delta N \equiv \Delta n/(\mathcal{P}/L)$, ξ in ΔP is replaced by η. The modified Bessel functions I_0 and I_1 apply for the upper sign in (144) and (146), that is,

for ν real, while the Bessel functions J_0 and J_1 apply for ν imaginary. The term in ΔP and ΔN with the delta function $\delta(U - X) = v_0\tau\delta(v_0t - x)$ represents a contribution that drifts at velocity v_0. The continuous distributions are confined to the interval $0 \leqq x \leqq v_0t$, $\mathbf{1}[X(U - X)] = \mathbf{1}[x(v_0t - x)]$ being the step function that is respectively zero and unity for negative and positive values of its argument. Note that κ and

$$\begin{aligned} \tfrac{1}{2}(\zeta + \kappa) &= \nu_v\tau, \\ \tfrac{1}{2}(\xi - \kappa) &= (1 - n_0/p_0)^{-1}(\nu_{tn1} - \nu_{tp1})\tau, \\ \tfrac{1}{2}(\eta - \kappa) &= (p_0/n_0 - 1)^{-1}(\nu_{tn1} - \nu_{tp1})\tau, \\ \tfrac{1}{2}(\xi - \eta) &= (\nu_{tn1} - \nu_{tp1})\tau \end{aligned} \tag{156}$$

are not restricted as to sign. For numerical computation of solutions and further analytical study, it is well to transform (155) by eliminating X in accordance with

$$X \equiv U \sin^2 \tfrac{1}{2}\Theta = \tfrac{1}{2}U(1 - \cos\Theta), \qquad U > 0, \tag{157}$$

which gives

$$\begin{aligned} \Delta P = e^{-\frac{1}{2}U(\zeta + \kappa\cos\Theta)}\Big\{ & [\tfrac{1}{2}U(\pi - \Theta)]^{-1}\delta(\pi - \Theta) \\ & + \tfrac{1}{2}[(\xi - \kappa)\begin{matrix} I_0 \\ J_0 \end{matrix}(\tfrac{1}{2}U\sin\Theta)\begin{matrix} +I_1 \\ -J_1 \end{matrix}(\tfrac{1}{2}U\sin\Theta)\tan\tfrac{1}{2}\Theta] \\ & \cdot\mathbf{1}[\Theta(\pi - \Theta)]\Big\} \end{aligned} \tag{158}$$

$$\begin{aligned} \Delta\hat{N} = \tfrac{1}{2}(\xi - \eta)[\, e^{-\frac{1}{2}U(\zeta + \kappa\cos\Theta)}] & \begin{matrix} I_0 \\ J_0 \end{matrix}(\tfrac{1}{2}U\sin\Theta) \\ & \cdot\mathbf{1}[\Theta(\pi - \Theta)]. \end{aligned}$$

For ΔN, ξ in ΔP is replaced by η. The use of Θ as a variable implies the step function of (155), while the step function of (158) simply restricts Θ as defined by (157) to the interval $0 \leqq \Theta \leqq \pi$.

Interpretation in descriptive terms of cases of imaginary ν requires further analysis. Illustrative cases of minority-carrier trapping in strongly extrinsic material, for which ν is real and whose interpretation is comparatively straightforward, will be presented first. For strongly extrinsic material, since the parameter ξ or η for minority carriers is substantially equal to κ, the minority-carrier concentration does not include the term with the Bessel function I_0. If, also, the trapping is nonrecombinative, then $\zeta = (\nu_g + \nu_t)\tau$ and $\kappa = (\nu_g - \nu_t)\tau$ hold with $\nu^2 = 4\nu_t\nu_g$, where ν_t and ν_g are ν_{tn1} or ν_{tp1} and ν_{gn1} or ν_{gp1}, respectively, and refer to the minority carrier. Figs. 1 and 2 show distributions of mobile minority

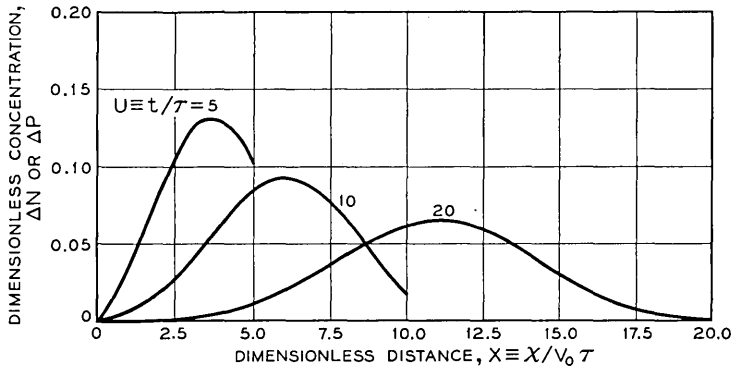


Fig. 1 — Continuous concentration distributions at different times of mobile minority carriers from an injected delta pulse for drift with trapping. A strongly extrinsic semiconductor and $\zeta = 1$, $\kappa = 0$ are assumed. For nonrecombinative trapping, these assumptions imply trapping time τ_t and release time τ_g both equal to twice the time unit τ. The pulse at the limit of the drift range is attenuated by the factor $e^{-\frac{1}{4}(\zeta-\kappa)U} = e^{-\frac{1}{4}U}$.

carriers for this case. For Fig. 1, ζ is unity and κ zero, as for trapping and release times equal,† and the continuous distributions are shown for different times after injection at the origin of the neutral delta pulse. These distributions are led by a delta pulse, which drifts at the ambipolar velocity v_0. This remnant of the initial pulse is composed of un-

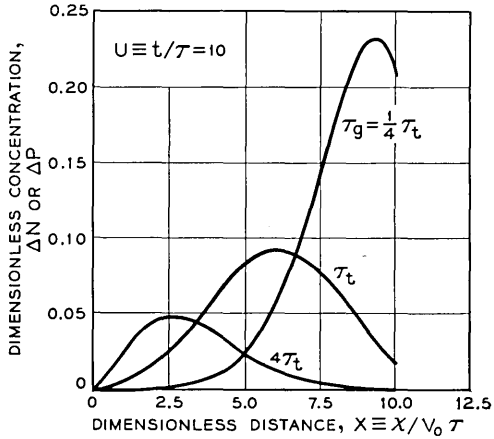


Fig. 2 — Continuous concentration distributions of mobile minority carriers from an injected delta pulse, for drift with trapping, all at time 10τ and for release time respectively 4, 1 and $\frac{1}{4}$ times trapping time. A strongly extrinsic semiconductor and nonrecombinative trapping are assumed; ζ equals $\frac{5}{4}$, 1, $\frac{5}{4}$; κ equals $-\frac{3}{4}$, 0, $\frac{3}{4}$; τ equals τ_t, $\frac{1}{2}\tau_t = \frac{1}{2}\tau_g$, τ_g, respectively.

† Fig. 1 applies more generally: The values of the parameters do not rule out recombination, but imply merely $\nu = \nu_s = 2\nu_{tn1} = 2[(\nu_{gn1}/\nu_{tn1})(\nu_{tn1}^2 - \nu_{tp1}^2)]^{\frac{1}{2}}$ for minority electrons and similar relations for minority holes.

trapped carriers and is rapidly attenuated by the exponential factor, which is $e^{-\frac{1}{2}U}$ for the particular values of the parameters. The abrupt fronts of the distributions for the shorter times result from most carriers having been trapped only slightly — at least once, but not much more. A relative maximum† appears in the case of the figure for times greater than $2\frac{1}{2}\tau$. For the longer times, the abrupt front disappears as a result of multiple trapping. Furthermore, there is a reduction of apparent mobility: The maximum ultimately drifts at velocity $\frac{1}{2}v_0$, the fraction of v_0 equal to the fraction of the time the carriers are free; this equality will be shown to apply for nonrecombinative minority-carrier trapping in strongly extrinsic material.[37] This limiting behavior sets in rather slowly, as the distribution for $t = 20\tau$ shows; its maximum occurs somewhat beyond the middle of the drift range.

Fig. 2 shows distributions all at time $t = 10\tau$ for release time respectively 4, 1 and $\frac{1}{4}$ times trapping time. The increasing areas under these distributions are associated with decreasing fractions of carriers trapped; it will be shown that, for nonrecombinative trapping, this trapped fraction rapidly approaches the fraction of the time carriers are trapped. The distributions have maxima appreciably beyond the respective values for large U of one-fifth, one-half and four-fifths of the drift range, and the distribution for the comparatively small release time still exhibits a high abrupt front at the time 10τ.

The parameters on which the solutions depend have certain general properties. From the first forms for κ and ζ of (145) and the definitions of τ and ν^2 of (143),

$$\zeta = (\kappa^2 \pm 1 + 4\Delta_1\tau^2)^{\frac{1}{2}} \geqq (\kappa^2 \pm 1)^{\frac{1}{2}} \tag{159}$$

follows. The inequality sign is associated with recombination, Δ_1 being zero for nonrecombinative trapping. The parameter ζ is real and never negative. For ν imaginary, so that the lower sign applies, a similar calculation gives

$$\kappa^2 = 1 + (\nu_s^2 - 4\Delta_1)\tau^2 \geqq 1, \qquad \nu^2 < 0; \tag{160}$$

the condition $\nu_s^2 - 4\Delta_1 \geqq 0$ implies real decay constants and holds from (85). For ν real, κ is not restricted. For example, for nonrecombinative trapping in strongly extrinsic material κ is $\frac{1}{2}[(\nu_g/\nu_t)^{\frac{1}{2}} - (\nu_t/\nu_g)^{\frac{1}{2}}]$ and can be zero (as for Fig. 1) or have any positive or negative value. Thus, $\zeta \geqq 1$ holds for ν real and $\zeta \geqq 0$ holds for ν imaginary. With (151) and

† Expressions for $d\Delta N/dX = (\frac{1}{2}U \sin \Theta)^{-1} d\Delta N/d\Theta$ from the Maclaurin's expansions for the modified Bessel functions reduce, for $\zeta = 1$ and $\kappa = \eta = 0$, to $\frac{1}{4}e^{-\frac{1}{2}U}$ and $\frac{1}{4}(1 - \frac{1}{8}U^2)\, e^{-\frac{1}{2}U}$ at the origin and at the end of the drift range.

(152), which define ξ and η, it is readily seen that $\xi \sim \eta \sim \kappa$ holds under the condition for the validity in general of the lifetime function $\tau_n \sim \tau_p$ of (75); and that ξ or η is approximately κ for $n_0 \gg p_0$ or $p_0 \gg n_0$, respectively. These properties are consistent with the easily verified relationship,

$$n_0\xi - p_0\eta = (n_0 - p_0)\kappa. \tag{161}$$

Also, for nonrecombinative trapping, the parameters do not depend on the capture coefficient and $\zeta = (\kappa^2 \pm 1)^{\frac{1}{2}}$ equals ξ or η according to whether electrons or holes are trapped.

Qualitative and physical distinctions are related principally to the signs of ν^2 and κ. The general condition for real ν is

$$(\nu_{tn1} + \nu_{tp1})(p_0 - n_0)/(\nu_{tn1} - \nu_{tp1}) > \mathfrak{N}_1^*, \tag{162}$$

from (143), and ν is imaginary if the inequality is reversed. From (54), the left-hand side may be written as $(p_0 + n_1^*)(p_0 - n_0)/(p_0 - n_1^*) = (p_1^* + n_0)(p_0 - n_0)/(p_1^* - n_0)$. The condition $\kappa > 0$ is, for p-type material,

$$p_0 - n_0 > (\nu_{tn1} - \nu_{tp1})(p_0 + n_0)/(\nu_{gn1} + \nu_{gp1}), \tag{163}$$

from the first equation of (145); changing signs of both sides or reversing the inequality gives $\kappa > 0$ for n-type or $\kappa < 0$ for p-type. From (54) and (63), the right-hand side may be written as

$$\frac{(1 + n_0/p_0)(p_0 - n_1^*)\mathfrak{N}_1^*}{n_0 + n_1^*} = \frac{(1 + p_0/n_0)(p_1^* - n_0)\mathfrak{N}_1^*}{p_1^* + p_0}.$$

If recombinative trapping is excluded, then ν is evidently real in the limit of strongly extrinsic material. Note, for example, that as p_0 increases indefinitely, ν_{tp1} approaches zero while ν_{tn1} approaches $C_{n1}\mathfrak{N}_1 = \tau_{n0}^{-1}$, so that ν approaches $2C_{n1}(n_1\mathfrak{N}_1)^{\frac{1}{2}}$. Also, since ν_{gp1} increases indefinitely and ν_{gn1} approaches $C_{n1}n_1$, κ becomes positively infinite. If ν_{tn1} and ν_{tp1} are equal for given p_0, then $\nu_{tn1} > \nu_{tp1}$ holds for all larger p_0. Suppose first that these capture frequencies are equal for some p_0 of the n-type material. Then, as p_0 decreases from a large value, κ does likewise. With $\nu_{tn1} > \nu_{tp1}$, (162) and (163) show that κ decreases to zero and becomes negative for ν still real. Further decrease of p_0 results in κ becoming negatively infinite, since ν approaches zero as $p_0 - n_0$ approaches $(\nu_{tn1} - \nu_{tp1})\mathfrak{N}_1^*/(\nu_{tn1} + \nu_{tp1})$, following which κ increases to -1 with ν imaginary. While ν is not defined for intrinsic material, the equations show that, in the approach to the intrinsic limit, ν is imaginary and κ approaches ± 1, the sign being that of $(\nu_{tn1} - \nu_{tp1})/(n_0 - p_0)$.

It is evident from (162) and (163) that, for n-type material and $\nu_{tn1} > \nu_{tp1}$ — that is, in the majority-carrier trapping range — ν is imaginary and κ positive, hence greater than unity. With increasing n_0, ν_{tn1} and ν_{tp1} ultimately approach equality, ν approaches zero and κ becomes positively infinite. For p-type material and $\nu_{tn1} > \nu_{tp1}$, or the minority-carrier capture range, it is likewise evident that positive κ implies real ν and that imaginary ν implies $\kappa < -1$. This latter case includes reverse drift. From a result of Section 2.2.1.2, since ν_{tn1} equals ν_{tp1} in n-type material, ν_{gn1} equals ν_{gp1} in p-type. Decrease of p_0 from the value for $\nu_{gn1} = \nu_{gp1}$ gives $\nu_{gn1} > \nu_{gp1}$, and, from (139), as p_0 approaches n_0, negative ν_v occurs. In the limit of negatively infinite κ, for which ν is zero (with $\nu_{tn1} \neq \nu_{tp1}$), ν_v is in general positive. Thus, the reverse-drift range is in general the portion of the minority-carrier capture range of imaginary ν that results if a certain infinite range of large negative values of κ is excluded. For nonrecombinative trapping, (139) and (162) both yield $| p_0 - n_0 | < \mathfrak{N}_1^*$, so that the two ranges coincide.

If the capture frequencies are equal for some p_0 of the p-type material, the initial decrease of κ as p_0 decreases from a large value still obtains; but the p_0 for equal capture frequencies is approached for ν still real and κ positive, and κ again becomes positively infinite as ν approaches zero. Imaginary ν results with further decrease of p_0 so that $\nu_{tp1} > \nu_{tn1}$ results, κ decreasing from large positive values to unity and then from -1 to large negative values, the majority-carrier trapping and minority-carrier capture ranges, respectively, being realized (for hole-capture frequency the larger) in p- and n-type material. It is easily shown, as before, that the reverse-drift range applies, with recombination, for a finite range of negative values of κ less than -1 in this minority-carrier capture range of imaginary ν. Increase of n_0 beyond the value for negatively infinite κ given (as before) by $n_0 - p_0 = (\nu_{tp1} - \nu_{tn1})/(\nu_{tp1} + \nu_{tn1})$ results in real ν, with which κ ranges from large negative values and becomes positively infinite as the material becomes strongly n-type.

For nonrecombinative trapping, κ for real ν and strongly extrinsic material is, as has been noted, a positive or negative constant. For electron trapping, for example, ν is $2C_{n1}(n_1\mathfrak{N}_1)^{\frac{1}{2}}$ and κ is $\frac{1}{2}[(n_1/\mathfrak{N}_1)^{\frac{1}{2}} - (\mathfrak{N}_1/n_1)^{\frac{1}{2}}]$ in the limit of large p_0; κ is positively infinite in the limit of large n_0. As p_0 decreases in the minority-carrier capture range of real ν, κ becomes negatively infinite as $p_0 - n_0$ approaches $\mathfrak{N}_1^*$. With further decrease of p_0, κ increases to -1 in the minority-carrier capture range of imaginary ν, which (for nonrecombinative trapping) is the reverse-drift range. In the majority-carrier trapping range in general, ν is imaginary and κ greater than unity.

Maxima for large U of each of the continuous distributions for cases of real ν occur substantially together. From (158), the Θ for a maximum is found to be given by† $\tan \Theta = -\kappa^{-1}$, which gives‡ $X/U = x/v_0 t = \frac{1}{2}[1 + \kappa/(\kappa^2 + 1)^{\frac{1}{2}}]$. For nonrecombinative traps, real ν implies minority-carrier trapping with positive ν_v and $(\kappa^2 + 1)^{\frac{1}{2}}$ equal to ζ. From (63), (136) and (145), X/U accordingly reduces to $\nu_v/(\nu_t + \nu_g)$, which is $[\tau_{tn1} - n_0\tau_{gn1}/(p_0 - n_0)]/(\tau_{tn1} + \tau_{gn1})$ for electron trapping or $[\tau_{tp1} - p_0\tau_{gp1}/(n_0 - p_0)]/(\tau_{tp1} + \tau_{gp1})$ for hole trapping. Hence X/U, the factor by which the apparent mobility is smaller than the magnitude of the ambipolar pseudomobility,[10] is in general less than $\tau_t/(\tau_t + \tau_g)$, the fraction of the time minority carriers are free; but X/U is substantially equal to this fraction[37] under the condition $|\, n_0 - p_0 \,| \gg \mathfrak{N}_1^*$, obtained by use of (63). As $|\, n_0 - p_0 \,|$ approaches $\mathfrak{N}_1^*$ in the nonrecombinative case, X/U approaches zero. Recombination reduces the distance for a maximum at given time, and thus reduces the apparent mobility, since, for nonrecombinative traps with recombination of lifetime τ_3 in other centers, the distribution of the mobile carriers subject to trapping is simply that for no recombination multiplied by the decay factor $e^{-x/v_0\tau_3}$. This factor applies because the carriers which arrive at x at whatever time have drifted in the conduction band for time x/v_0.

The decay constant for the straggle effect[54] is that of the limiting decay of the tail of the distribution at fixed x after the maximum has passed. It is given by the exponent in (158), and is accordingly ν_v. This result follows from (156), since, from (157), $x \ll v_0 t$ implies $\cos \Theta \sim 1$. Real ν and hence positive ν_v obtain in this connection. By use of (63), it is easily shown that ν_v for strongly extrinsic material is substantially $\nu_{gn1} + \nu_{gp1}$ plus either ν_{tn1}, for $n_0 \gg p_0$, or ν_{tp1}, for $p_0 \gg n_0$.

Integrals of the solutions of (158) over the drift range are evaluated in Appendix C. These integrals give

$$\begin{aligned} F_p &\equiv \mathcal{P}^{-1} \int_0^{v_0 t} \Delta p \, dx = \int_0^U \Delta P \, dX \\ &= e^{-\frac{1}{2}\zeta U}[\xi(\kappa^2 \pm 1)^{-\frac{1}{2}} \sinh \tfrac{1}{2}(\kappa^2 \pm 1)^{\frac{1}{2}} U + \cosh \tfrac{1}{2}(\kappa^2 \pm 1)^{\frac{1}{2}} U], \\ \hat{F}_n &\equiv \mathcal{P}^{-1} \int_0^{v_0 t} \Delta \hat{n} \, dx = \int_0^U \Delta \hat{N} \, dX \\ &= e^{-\frac{1}{2}\zeta U}(\xi - \eta)(\kappa^2 \pm 1)^{-\frac{1}{2}} \sinh \tfrac{1}{2}(\kappa^2 \pm 1)^{\frac{1}{2}} U \end{aligned} \tag{164}$$

† Use is made of the approximations $I_0(z) \sim I_1(z) \sim (2\pi z)^{-\frac{1}{2}} e^z$ for $|z|$ large. The distributions for large U are substantially proportional and gaussian in shape. For nonrecombinative minority-carrier trapping, they are as if the excess majority carriers were subject only to drift and diffusion with diffusivity $v_0 L/4\zeta^3$.

‡ Note that, for the maximum, $\Theta - \pi/2$ has the sign of κ, so that $\cos \Theta$ has the opposite sign.

as mobile and trapped fractions at given time of carriers initially injected. For

$$F_n \equiv \mathcal{P}^{-1} \int_0^{v_0 t} \Delta n \, dx = F_p - \hat{F}_n ,$$

ξ in F_p is replaced by η. Equation (159) serves to verify that, with recombination, F_p , $\hat{F}_n$ and F_n all approach zero as U increases indefinitely. For nonrecombinative electron trapping, $\xi = \zeta = (\kappa^2 \pm 1)^{\frac{1}{2}}$ gives $F_p = 1$, as may be expected, and $\hat{F}_n = \frac{1}{2}(1 - \eta/\zeta)(1 - e^{-\zeta U})$. Thus, from (145) and (152), the trapped fraction approaches $\tau_{gn1}/(\tau_{tn1} + \tau_{gn1})$, the fraction of the time electrons are trapped, with time constant $\tau_{gn1}\tau_{tn1}/(\tau_{tn1} + \tau_{gn1})$; the mobile fraction approaches the fraction of the time electrons are free. For hole trapping, entirely similar results apply. All of these results evidently apply for ν imaginary as well as real.

In Fig. 3 are shown continuous minority-carrier distributions for imaginary ν, in particular, for nonrecombinative trapping of minority carriers in the reverse-drift range and also of majority carriers. In the former case, an attenuated delta pulse of untrapped mobile carriers, which drifts at velocity v_0 , leads a continuous distribution of minority carriers that crowds towards the origin as its maximum excursions above and below the axis both increase with time. The distribution of added minority carriers is negative over part of the drift range after a certain time.† In accordance with (164), it approaches a net positive delta pulse at the origin of strength $\mathcal{P}$, the initial strength, times the free-time fraction. In the case of majority-carrier trapping, the pulse of untrapped carriers increases in strength as it drifts at velocity v_0 . This augmentation is appreciable with appreciable equilibrium minority-carrier concentration; it is negligible in strongly extrinsic material. The pulse leads a largely or entirely negative continuous distribution of minority carriers, which crowds towards the pulse as its excursion below the axis increases with time.‡ This distribution approaches a negative delta pulse, which, with the minority carriers of the augmented pulse, gives a net pulse of strength $\mathcal{P}$. The corresponding distributions of mobile and trapped majority carriers approach net pulses of strengths equal to $\mathcal{P}$ times the free- and trapped-time fractions, respectively.

These results exhibit, with due allowance for the neglect of diffusion,

† Negative concentrations of the trapped and majority carriers also occur. For this case, $\eta - \kappa$ is positive for p-type material, and, if it is not too large, negative Δn first appears (as for the case of the figure) at the end of the drift range for times greater than $2(\eta - \kappa)\tau$, since Δn equals $\frac{1}{2}[e^{-\frac{1}{2}(\zeta-\kappa)}]^U(\eta - \kappa - \frac{1}{2}U)$ for $\Theta = \pi$. With sufficiently large $\eta - \kappa$, negative Δn first appears within the range.

‡ The majority carriers are similarly distributed: For the case of the figure, (negative) dimensionless majority-carrier concentration ΔP or ΔN is $\frac{3}{2}\Delta N$ or $\frac{3}{2}\Delta P$ for $X = U = 5$ and $\frac{14}{11}\Delta N$ or $\frac{14}{11}\Delta P$ for $X = U = 10$.

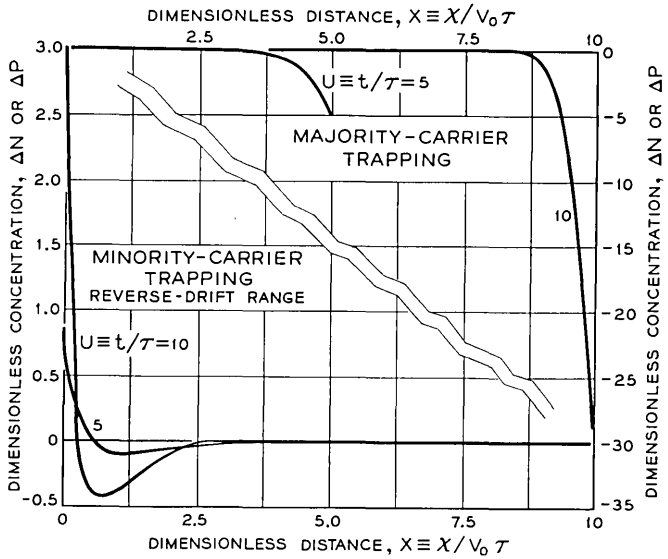


Fig. 3 — Continuous concentration distributions at different times of mobile minority carriers from an injected delta pulse, for drift with minority-carrier trapping in the reverse-drift range and for drift with majority-carrier trapping. Equilibrium majority-carrier concentration is taken as $2n_i$ for both cases, and nonrecombinative trapping is assumed, with τ_g/τ_t large compared with the equilibrium ratio of majority-carrier concentration excess to concentration of the (mobile) carriers that are subject to trapping. For minority-carrier trapping, a strong inequality for reverse drift accordingly holds, with $\tau_g \gg 3\tau_t = 4\tau$, $\zeta = \frac{3}{4}$, $\kappa = -\frac{5}{4}$, and attenuation of the pulse by the factor e^{-U}. For majority-carrier trapping, $\tau_g \gg \frac{3}{4}\tau_t = \tau$, $\zeta = \frac{3}{4}$, $\kappa = \frac{5}{4}$ hold, and the pulse is augmented by the factor $e^{\frac{1}{4}U}$. For both cases, $\xi = -\eta = \frac{3}{4}$ or $\eta = -\xi = \frac{3}{4}$ holds for electrons or holes trapped, so that the coefficients for the minority carriers of the terms in J_0 for the two cases are respectively $\frac{1}{4}$ and $-\frac{1}{4}$.

the essential behavior that would be realized in practice. In each case, the initial neutral pulse of mobile carriers will appear essentially as a diffuse pulse of trapped and mobile carriers in substantially the proportions that obtain for the steady state. Minority-carrier trapping in the reverse-drift range does not give largely unidirectional drift, while majority-carrier trapping results essentially in drift at the ambipolar velocity v_0. Increase of applied field, however, makes for the idealized behavior shown in the figure—in particular, for continuous distributions that are negative over increased actual distance at given time. Numerical estimate of the effect of diffusion in the reverse-drift case shows that negative added-carrier concentrations can occur over appreciable distances under conditions that can be realized in practice for reasonable values of trapping time.†

† For Fig. 3, $\frac{1}{4}\tau_{gn1} \gg \tau = \frac{3}{4}\tau_{tn1}$ holds for p-type material, and if τ is 10^{-5} second, then $L = v_0\tau$ is about 0.4 cm for $p_0 = 2n_i$ and electron and hole mobilities of 1500

A simple descriptive interpretation of the crowding of the distributions towards the origin in the case of reverse drift is that the drift of added carriers, initially in the direction of v_0, is largely in the opposite direction, the direction of drift of majority carriers, after some trapping has taken place. For more detailed general interpretations, consider the current density $\Delta \mathbf{I}$ of added carriers of (19), which is given under the present assumptions by

$$\begin{aligned}\Delta I &= e^2\mu_n\mu_p\sigma_0^{-2}I(\mathcal{P}/L)(n_0\Delta P - p_0\Delta N)\\ &= ev_0(\mathcal{P}/L)[e^{-\frac{1}{2}U(\zeta+\kappa\cos\Theta)}\Big\{[\tfrac{1}{2}U(\pi-\Theta)]^{-1}\delta(\pi-\Theta)\\ &\cdot \begin{matrix}+\frac{1}{2}I_1\\ -\frac{1}{2}J_1\end{matrix}(\tfrac{1}{2}U\sin\Theta)\tan\tfrac{1}{2}\Theta\,\mathbf{1}[\Theta(\pi-\Theta)]\Big\},\end{aligned} \tag{165}$$

from (8), (158) and (161). With (158), this result shows that the Bessel functions of order zero are associated with carriers that neutralize the charge of trapped carriers or with the trapped carriers themselves, while those of order one are associated with the drift of, in effect, carrier pairs. The direction of drift of a mobile-carrier distribution considered in its entirety depends on the sign of the net ΔI, or ΔI integrated over the drift range. By use of (161) and (164),

$$\begin{aligned}\int_0^U \Delta I dX &= e^2\mu_n\mu_p\sigma_0^{-2}I(\mathcal{P}/L)(n_0F_p - p_0F_n)\\ &= ev_0(\mathcal{P}/L)(e^{-\frac{1}{2}\kappa U})[\kappa(\kappa^2\pm 1)^{-\frac{1}{2}}\sinh\tfrac{1}{2}(\kappa^2\pm 1)^{\frac{1}{2}}U\\ &\qquad + \cosh\tfrac{1}{2}(\kappa^2\pm 1)^{\frac{1}{2}}U]\end{aligned} \tag{166}$$

results. For nonrecombinative electron trapping, this integral reduces to $ev_0(\mathcal{P}/L)$, the initial ΔI, times† $\frac{1}{2}(1+\kappa/\zeta) = \nu_v/(\nu_{tn1}+\nu_{gn1})$ in the limit for U infinite, from (136), (145) and (159). As may be expected, the limiting integral has the sign of v_0 or the opposite sign according to whether ν_v is positive or negative. That the distributions ultimately crowd towards the origin in the case of reverse drift is established on a more quantitative basis by this result.

---

and 570 cm^2 volt^{-1} second^{-1} (as for silicon[72,73] at 300°K) and for an applied field of 100 volt cm^{-1}, which is reasonable for a filament of resistivity at least a few hundred ohm-cm and of area of cross section about 10^{-2} cm or less. The corresponding diffusion distance $(D_0t)^{\frac{1}{2}}$ for $t = 10\tau$ is 0.054 cm, an order of magnitude smaller than the approximate distance $2L$ of 0.8 cm over which negative added-carrier concentrations occur. This difference is greater for larger τ_{tn1}.

† For real ν, which implies positive ν_v and p-type material, this factor is the limiting value of X/U for the maximum of the mobile-electron distribution.

It follows from (165) and is otherwise evident that ΔI is zero at the origin.† Also, for imaginary ν, the Bessel function with the minus sign applies and ΔI within the drift range is negative for positive v_0 if U is not too large. Hence, in the case of reverse drift, the mobile-carrier distributions become steeper even for small times.‡ This conclusion is consistent with the increase with U of the exponential factor for given small Θ, for which $\frac{1}{2}(\zeta + \kappa \cos \Theta) \sim \nu_v \tau$ is negative, and also with the decrease with increasing Θ of this factor for given U, which results because of $\kappa < -1$. With electron trapping, Δn becomes small for sufficiently large U at some Θ. But, with $\xi > \eta$, since $\Delta p \sim \Delta \hat{n}$ is still positive at this Θ, ΔI is still negative, from (165), I being negative for positive v_0. Thus, Δn becomes negative. As Δn becomes increasingly negative, electrons tend to be released from traps and $\Delta \hat{n}$ in turn becomes small [as a zero of $J_0(\frac{1}{2}U \sin \Theta)$ is approached]; but ΔI is still negative. Thus, $\Delta \hat{n}$ becomes negative too. To the right of the location at which $\Delta \hat{n}$ equals $(p_0/n_0 - 1)\Delta n < 0$, which is the condition $\Delta I = 0$ (for negative Δn, Δp and $\Delta \hat{n}$), ΔI is positive and the concentrations tend to increase algebraically. A progressive increase with time of the maximum excursion of Δn below the axis is associated with the presence of a (time-dependent) location at which ΔI changes sign.

In the case of majority-carrier trapping, the exponential factor for given U increases with Θ because of $\kappa > 1$; and $\zeta - \kappa$ being negative,§ this factor is smaller or greater than unity respectively for Θ small or $\Theta \sim \pi$. Thus, as U increases, the continuous minority-carrier distribution, which is negative for all Θ and U, crowds towards the pulse, which drifts at velocity v_0, and both increase in amplitude. For electron trapping in general, $\Delta \hat{n} > 0$ holds at (and near) the pulse. If the material is sufficiently strongly n-type, this $\Delta \hat{n}$ is largely compensated entirely by negative Δn. The pulse, in effect, removes electrons from the semiconductor and transfers them to traps, and its strength remains substantially constant; the continuous contribution to ΔI is negligible because of large τ. In less strongly n-type material, however, neutrality is maintained at the pulse in part by negative Δp and correspondingly more negative Δn, which makes for less trapping for given pulse strength. But holes removed from the semiconductor drift at velocity v_0 and thus (with an equal number of electrons) cause a progressive increase in the strength of the pulse; a $\Delta \hat{n}$ obtains that does not depend explicitly on n_0. The ΔI of the pulse likewise increases indefinitely. The net or inte-

† Thus, at the origin, $\Delta n/\Delta p$ maintains the value n_0/p_0.

‡ Note also that (158) gives, for small $U > 0$, ΔP or ΔN at the origin equal to $\xi - \kappa$ or $\eta - \kappa$ times $(1 - \nu_v t)$, which increases with t in the case of reverse drift.

§ Note that (159) may be written as $(\zeta + \kappa)(\zeta - \kappa) = -1$ for nonrecombinative trapping and imaginary ν.

grated ΔI, however, increases to a limiting value, as given by the factor $\nu_v/(\nu_{tn1} + \nu_{gn1})$, which is greater than unity in the present case. The largest excursions of the continuous mobile-carrier distributions below the axis occur, of course, at the pulse, at which there is a discontinuity in ΔI with change of sign.

In cases of real ν, the exponential factor is less than unity and decreases as U increases.† With the monotonically increasing modified Bessel functions and ΔI everywhere in the direction of v_0, the continuous distributions, and, from (165), ΔI as well, ultimately possess relative maxima that drift in this direction with a common velocity. While the maxima occur substantially together, the distribution of the trapped minority carriers lags, consistent with the Bessel function of order zero dominating that of order one; the mobile minority-carrier distribution ultimately results entirely from carriers released from traps.

The case of "critical trapping," the borderline case between cases of real and imaginary ν, is one whose analysis furnishes further qualitative insight. For it, ν as defined in (143) is zero with $\nu_{tn1} \neq \nu_{tp1}$, and the condition is the equality corresponding to (162). The previous notation can be used for this case by choosing the time unit τ arbitrarily and noting that (144) through (154) then all apply if the terms with the double signs are omitted wherever they occur. A calculation essentially similar to that given in Appendix B provides the solution, which, for the initial delta pulse of $\mathcal{P}$ carrier pairs per unit area at the origin, is

$$\begin{aligned}
\Delta p &= (\mathcal{P}/L)\{\exp\,[-\tfrac{1}{2}(\zeta - \kappa)U]\cdot\delta(U - X) \\
&\quad + \tfrac{1}{2}(\xi - \kappa)\,\exp\,[\kappa X - \tfrac{1}{2}(\zeta + \kappa)U]\cdot\mathbf{1}[X(U - X)]\} \\
&= \mathcal{P}\{\exp\,[-\tfrac{1}{2}(\nu_{tn1} + \nu_{tp1})(1 + (n_0 + p_0)/\mathfrak{N}_1{}^*)t]\cdot\delta(v_0t - x) \quad (167) \\
&\quad + [p_0(\nu_{tn1} + \nu_{tp1})/v_0\mathfrak{N}_1{}^*]\,\exp\,[-(v_0\mathfrak{N}_1{}^*)^{-1} \\
&\quad\quad \cdot(p_0\nu_{tn1} + n_0\nu_{tp1})x - t/\tfrac{1}{2}(\tau_{tn1} + \tau_{tp1})]\cdot\mathbf{1}[x(v_0t - x)]\}, \\
\Delta\hat{n} &= \tfrac{1}{2}(\mathcal{P}/L)(\xi - \eta)\,\exp\,[\kappa X - \tfrac{1}{2}(\kappa + \zeta)U]\cdot\mathbf{1}[X(U - X)] \\
&= \mathcal{P}[(p_0 - n_0)(\nu_{tn1} + \nu_{tp1})/v_0\mathfrak{N}_1{}^*]\,\exp\,[-(v_0\mathfrak{N}_1{}^*)^{-1} \\
&\quad\quad \cdot(p_0\nu_{tn1} + n_0\nu_{tp1})x - t/\tfrac{1}{2}(\tau_{tn1} + \tau_{tp1})]\cdot\mathbf{1}[x(v_0t - x)].
\end{aligned}$$

For Δn, ξ is replaced by η in the first expression for Δp, or n_0 and p_0 as

† In these cases, $\zeta + \kappa$ has the sign of ν_v and is positive, from (156), and (159) may be written for nonrecombinative trapping as $(\zeta + \kappa)(\zeta - \kappa) = 1$.

well as subscripts n and p are interchanged in the second. The concentration increments in terms of the dimensionless quantities are, of course, actually independent of τ. In the second expressions, quantities involved have been written in forms that apply under the condition for critical trapping. It follows readily from these equations that the mobile and trapped fractions are similar to those of (164) except that $(\kappa^2 \pm 1)^{\frac{1}{2}}$ is replaced by κ; they are independent of τ. For nonrecombinative trapping, the mobile and trapped fractions of carriers respectively approach, as before, the free- and trapped-time fractions, since κ equals $(-\zeta)$.†

As the second expression for $\Delta\hat{n}$ shows, critical trapping is a case of minority-carrier trapping. The continuous distributions are proportional and are equal to products of exponential decays with distance and with time. The amplitudes of the distributions at the origin are larger and the decay with distance is sharper the smaller is v_0. The time decay results from recombination, with which neither τ_{tn1} nor τ_{tp1} is infinite. For nonrecombinative trapping, the condition of critical trapping is the same as that of zero drift, and exponential distributions are established progressively over the variable range $v_0 t$ that otherwise do not change with time. It is well to note that the general case of zero ν_v is a case of imaginary ν, with ν^2 equal to $(-4\Delta_1)$. Furthermore, for trapping in intrinsic material, for which ν_v is not defined, it can be shown that the distributions are all identically zero except at the origin. As may be expected for this case of no diffusion, the initial pulse results simply in pulses for the concentration increments that change as trapping and recombination proceed.

IV. ACKNOWLEDGMENTS

The author is indebted to J. A. Burton and to W. H. Brattain for discussions to which this work owes its inception, and is pleased to acknowledge their helpful interest and that of many others, among whom are included G. Bemski, W. L. Brown, T. M. Buck, R. J. Collins, C. G. B. Garrett, G. W. Gobeli, J. R. Haynes, J. J. Hopfield, H. J. Hrostowski, W. Kaiser, M. Lax, C. A. Lee, F. J. Morin, W. T. Read, R. G. Shulman, W. G. Spitzer, G. H. Wannier, G. K. Wertheim and P. A. Wolff. Some were of assistance in particular connections, as was H. Y. Fan, with a communication concerning his treatment of drift with trapping, and H. O. Pollak, who provided a method for evaluating the integral over the drift range of mobile-carrier concentration.

† Note that $\zeta = (\kappa^2 + 4\Delta_1)^{\frac{1}{2}}$ and $\kappa < 0$ hold.

APPENDIX A

Derivation of Two-Sided Laplace Transforms

The transforms

$$\mathfrak{L}\left\{\int_X^\infty \begin{matrix} I_0 \\ J_0 \end{matrix} [(\beta^2 - X^2)^{\frac{1}{2}}] f(\beta)\, d\beta\right\} = (s^2 \pm 1)^{-\frac{1}{2}} F[-(s^2 \pm 1)^{\frac{1}{2}}],$$

and (168)

$$\mathfrak{L}\left\{f(X) \pm \int_X^\infty \beta(\beta^2 - X^2)^{-\frac{1}{2}} \begin{matrix} I_1 \\ J_1 \end{matrix} [(\beta^2 - X^2)^{\frac{1}{2}}] f(\beta) d\beta\right\} = F[-(s^2 \pm 1)^{\frac{1}{2}}]$$

entail restrictions on $f(X)$ and on the transform variable s for convergence of the integrals. In accordance with the definition of (141), the first transform is

$$\begin{aligned}
&\int_{-\infty}^\infty e^{-s\lambda} \int_\lambda^\infty \begin{matrix} I_0 \\ J_0 \end{matrix} [(\beta^2 - \lambda^2)^{\frac{1}{2}}] f(\beta)\, d\beta\, d\lambda \\
&\quad = \int_{-\infty}^\infty f(\beta) \int_{-\infty}^\beta \begin{matrix} I_0 \\ J_0 \end{matrix} [(\beta^2 - \lambda^2)^{\frac{1}{2}}]\, e^{-s\lambda}\, d\lambda\, d\beta \\
&\quad = \int_0^\infty [f(\beta) + f(-\beta)] \int_\beta^\infty \begin{matrix} J_0 \\ I_0 \end{matrix} [(\lambda^2 - \beta^2)^{\frac{1}{2}}]\, e^{s\lambda}\, d\lambda\, d\beta \\
&\qquad + \int_0^\infty f(\beta) \int_{-\beta}^\beta \begin{matrix} J_0 \\ I_0 \end{matrix} [(\lambda^2 - \beta^2)^{\frac{1}{2}}]\, e^{-s\lambda}\, d\lambda\, d\beta,
\end{aligned} \tag{169}$$

as obtained by changing the order of integration and changing from variable β to $-\beta$ in the contribution from the range of negative β. The first of the integrals with respect to λ is†

$$\begin{aligned}
&\int_\beta^\infty \begin{matrix} J_0 \\ I_0 \end{matrix} [(\lambda^2 - \beta^2)^{\frac{1}{2}}]\, e^{s\lambda}\, d\lambda \\
&\quad = \int_0^\infty \mu(\mu^2 + \beta^2)^{-\frac{1}{2}} \begin{matrix} J_0 \\ I_0 \end{matrix} (\mu) \exp [s(\mu^2 + \beta^2)^{\frac{1}{2}}]\, d\mu \\
&\quad = (s^2 \pm 1)^{-\frac{1}{2}} \exp [-(s^2 \pm 1)^{\frac{1}{2}}\beta].
\end{aligned} \tag{170}$$

For convergence with the upper function and sign, s is taken as (real

† Watson,[93] p. 416, Equation (2).

and) negative; for convergence with the lower function and sign, the real part of s is to be less than -1. The second is a Gegenbauer integral: With

$$\begin{aligned} \lambda &\equiv -\beta \cos \phi, \\ z \sin \psi &\equiv \begin{pmatrix} i \\ 1 \end{pmatrix} \beta, \\ z \cos \psi &\equiv -is\beta, \\ z &= i(s^2 \pm 1)^{\frac{1}{2}}\beta, \\ \cot \psi &= -\begin{pmatrix} 1 \\ i \end{pmatrix} s, \end{aligned} \tag{171}$$

the result

$$\begin{aligned} &\int_{-\beta}^{\beta} \frac{J_0}{I_0} [(\lambda^2 - \beta^2)^{\frac{1}{2}}] e^{-s\lambda}\, d\lambda \\ &\qquad = \beta \int_0^{\pi} J_0(z \sin \phi \sin \psi)\, e^{iz \cos \phi \cos \psi} \sin \phi \, d\phi \\ &\qquad = 2(s^2 \pm 1)^{-\frac{1}{2}} \sinh [(s^2 \pm 1)^{\frac{1}{2}}\beta] \end{aligned} \tag{172}$$

follows.† By substituting from (170) and (172) in (169), the first transform of (168) is established. The second transform is established through

$$\begin{aligned} &\Re \left\{ -\int_X^{\infty} \frac{I_0}{J_0} [(\beta^2 - X^2)^{\frac{1}{2}}] f'(\beta)\, d\beta \right\} \\ &\qquad = \int_{-\infty}^{\infty} f(\lambda) \exp [(s^2 \pm 1)^{\frac{1}{2}}\lambda]\, d\lambda \\ &\qquad = F[-(s^2 \pm 1)^{\frac{1}{2}}]. \end{aligned} \tag{173}$$

The original expression reduces to the form on the left upon integrating by parts for $f(\beta)$ such that $I_0(\beta)f(\beta)$ or $J_0(\beta)f(\beta)$ is zero for $\beta = \infty$. The result then follows by application of the first transform and another integration by parts, which entails $f(\lambda) \exp [(s^2 \pm 1)^{\frac{1}{2}}\lambda]$ equal to zero for positively and negatively infinite λ, a condition that subsumes the preceding one.

† Watson[93] p. 379, Equation (1).

APPENDIX B

Solutions for Drift of an Injected Pulse

The dimensionless concentration $\Delta P \equiv \Delta p/(\mathcal{P}/L)$ is the inverse Laplace transform of $\overline{\Delta P}$ given by (154). It may be evaluated by use of

$$\mathfrak{L}^{-1} \exp\,(a^2s^2 - \tfrac{1}{2}Us) = \tfrac{1}{2}\pi^{-\frac{1}{2}}a^{-1} \exp\,[-\,(X - \tfrac{1}{2}U)^2/4a^2] \quad (174)$$

and the similar formula obtained by replacing U by its negative in conjunction with (168) and certain elementary rules. Equations (168) and (174) give

$$\mathfrak{L}^{-1}(s^2 \pm 1)^{-\frac{1}{2}} \exp\,[a^2(s^2 \pm 1) + \tfrac{1}{2}U(s^2 \pm 1)^{\frac{1}{2}}]$$

$$= \tfrac{1}{2}\pi^{-\frac{1}{2}}a^{-1} \int_X^\infty \frac{I_0}{J_0}\,[(\beta^2 - X^2)^{\frac{1}{2}}] \exp\,[-(\beta - \tfrac{1}{2}U)^2/4a^2]\,d\beta,$$

$$\mathfrak{L}^{-1} \exp\,[a^2(s^2 \pm 1) + \tfrac{1}{2}U(s^2 \pm 1)^{\frac{1}{2}}] \quad (175)$$

$$= \tfrac{1}{2}\pi^{-\frac{1}{2}}a^{-1} \left\{ \exp\,[-\,(X - \tfrac{1}{2}U)^2/4a^2] \pm \int_X^\infty \beta(\beta^2 - X^2)^{-\frac{1}{2}} \cdot \frac{I_1}{J_1}\,[(\beta^2 - X^2)^{\frac{1}{2}}] \exp\,[-(\beta - \tfrac{1}{2}U)^2/4a^2]\,d\beta \right\}.$$

The exponents that occur in (154) may be transformed in accordance with

$$\begin{aligned} a^2s^2 - \tfrac{1}{2}U(s + \zeta) + \tfrac{1}{2}U[(s - \kappa)^2 \pm 1]^{\frac{1}{2}} \\ \equiv a^2(\kappa^2 \mp 1) - \tfrac{1}{2}(\kappa + \zeta)U + (2a^2\kappa - \tfrac{1}{2}U)(s - \kappa) \\ + a^2[(s - \kappa)^2 \pm 1] + \tfrac{1}{2}U[(s - \kappa)^2 \pm 1]^{\frac{1}{2}} \end{aligned} \quad (176)$$

and the similar identity that holds with minus signs before the radicals. Consideration of the transformations that change the exponent in the transforms of (175) to that given by (176) shows that the inverse transforms or solutions sought contain $\exp\,[a^2(\kappa^2 \mp 1) - \tfrac{1}{2}(\kappa + \zeta)U + \kappa X]$ as factor, in which $\exp\,(\kappa X)$ results from replacing s by $s - \kappa$. They contain also $X + 2a^2\kappa - \tfrac{1}{2}U$ in place of X in (175), because of the term $(2a^2\kappa - \tfrac{1}{2}U)(s - \kappa)$ in (176). The solutions are obtained by straightforward application of these results to (154); corresponding to the s in the factor $s - \xi$ of the second terms, there are contributions obtained

by differentiating with respect to X the expressions that originate from the first equation of (175). The solutions are:

$$\begin{aligned}\Delta P = {} & \tfrac{1}{2}\pi^{-\frac{1}{2}}a^{-1}\{\exp[a^2(\kappa^2 \mp 1) - \tfrac{1}{2}(\zeta + \kappa)U + \kappa X]\} \\ & \cdot\Big(\exp[-(X + 2a^2\kappa - U)^2/4a^2] - \tfrac{1}{2}(\xi - \kappa) \\ & \cdot\int_{X+2a^2\kappa-\frac{1}{2}U}^{\infty} \frac{I_0}{J_0}\{[\beta^2 - (X + 2a^2\kappa - \tfrac{1}{2}U)^2]^{\frac{1}{2}}\} \\ & \cdot\{\exp[-(\beta + \tfrac{1}{2}U)^2/4a^2] - \exp[-(\beta - \tfrac{1}{2}U)^2/4a^2]\}\, d\beta \\ & \mp \frac{1}{2}\int_{X+2a^2\kappa-\frac{1}{2}U}^{\infty} [\beta^2 - (X + 2a^2\kappa - \tfrac{1}{2}U)^2]^{-\frac{1}{2}} \\ & \cdot\frac{I_1}{J_1}\{[\beta^2 - (X + 2a^2\kappa - \tfrac{1}{2}U)^2]^{\frac{1}{2}}\} \\ & \cdot\{(X + 2a^2\kappa - \tfrac{1}{2}U - \beta)\exp[-(\beta + \tfrac{1}{2}U)^2/4a^2] \\ & - (X + 2a^2\kappa - \tfrac{1}{2}U + \beta)\exp[-(\beta - \tfrac{1}{2}U)^2/4a^2]\}\, d\beta\Big).\end{aligned} \tag{177}$$

Replacing ξ by η in (177) gives the solutions for ΔN, and those for $\Delta\hat{N}$ are accordingly

$$\begin{aligned}\Delta\hat{N} = {} & \tfrac{1}{4}\pi^{-\frac{1}{2}}a^{-1}(\eta - \xi)\{\exp[a^2(\kappa^2 \mp 1) - \tfrac{1}{2}(\zeta + \kappa)U + \kappa X]\} \\ & \cdot\int_{X+2a^2\kappa-\frac{1}{2}U}^{\infty} \frac{I_0}{J_0}\{[\beta^2 - (X + 2a^2\kappa - \tfrac{1}{2}U)^2]^{\frac{1}{2}}\} \\ & \cdot\{\exp[-(\beta + \tfrac{1}{2}U)^2/4a^2] - \exp[-(\beta - \tfrac{1}{2}U)^2/4a^2]\}\, d\beta.\end{aligned} \tag{178}$$

The corresponding limiting solutions of (155) involve the step-function factor as result of the requirement that, for contributions in the limit of zero a, the gaussian factors in the integrands of (177) and (178) be centered at values within the range of integration.

APPENDIX C

Integrals Over the Drift Range

Consider first evaluation of $\hat{F}_n$, which may be written as

$$\begin{aligned}\hat{F}_n = {} & \tfrac{1}{4}(\xi - \eta)U[\exp(-\tfrac{1}{2}\zeta U)] \\ & \cdot\int_0^{\pi} [\exp(-\tfrac{1}{2}\kappa U\cos\Theta)]\frac{I_0}{J_0}(\tfrac{1}{2}U\sin\Theta)\sin\Theta\, d\Theta\end{aligned} \tag{179}$$

from (157) and (158). The transformations

$$\begin{aligned} z \sin \psi &\equiv \begin{pmatrix} i\tfrac{1}{2}U \\ \tfrac{1}{2}U \end{pmatrix}, \\ z \cos \psi &\equiv \begin{pmatrix} i\tfrac{1}{2}\kappa U \\ i\tfrac{1}{2}\kappa U \end{pmatrix}, \\ z &= i\tfrac{1}{2}(\kappa^2 \pm 1)^{\frac{1}{2}}U, \\ \cot \psi &= \begin{pmatrix} \kappa \\ i\kappa \end{pmatrix} \end{aligned} \tag{180}$$

result in the single form,

$$\hat{F}_n = \tfrac{1}{4}(\xi - \eta)U[\exp(-\tfrac{1}{2}\zeta U)] \int_0^\pi [\exp(iz \cos \Theta \cos \psi)] \cdot J_0(z \sin \Theta \sin \psi) \sin \Theta \, d\Theta; \tag{181}$$

the Gegenbauer's integral† reduces to $(2\pi/z)^{\frac{1}{2}}J_{\frac{1}{2}}(z) = 2z^{-1} \sin z$, and, with the definitions of z, the respective expressions for $\hat{F}_n$ of (164) follow.

The contribution to F_p from the Bessel function of order zero is clearly $[(\xi - \kappa)/(\xi - \eta)]\hat{F}_n$, while the contribution from the delta pulse at the limit of the drift range is $\exp[\tfrac{1}{2}(\kappa - \zeta)U]$. The contribution

$$\begin{aligned} &\pm\tfrac{1}{2}\{\exp[-\tfrac{1}{2}(\zeta + \kappa)U]\} \\ &\quad \cdot \int_0^U \lambda[\lambda(U - \lambda)]^{-\frac{1}{2}}[\exp(\kappa\lambda)] \begin{matrix} I_1 \\ J_1 \end{matrix} [(\lambda(U - \lambda))^{\frac{1}{2}}] \, d\lambda \\ &= \pm\tfrac{1}{4}U[\exp(-\tfrac{1}{2}\zeta U)] \int_0^\pi [\exp(-\tfrac{1}{2}\kappa U \cos \Theta)] \\ &\qquad \cdot \begin{matrix} I_1 \\ J_1 \end{matrix}(\tfrac{1}{2}U \sin \Theta)(1 - \cos \Theta) \, d\Theta \\ &= \begin{pmatrix} -i \\ -1 \end{pmatrix} \tfrac{1}{4}U[\exp(-\tfrac{1}{2}\zeta U)] \\ &\quad \cdot \int_0^\pi [\exp(iz \cos \Theta \cos \psi)]J_1(z \sin \Theta \sin \psi)(1 - \cos \Theta) \, d\Theta, \end{aligned} \tag{182}$$

the last form for which follows from (180), remains to be evaluated. A method due to Pollak[94] involves first writing the integral in the last form, to be denoted by $\mathcal{J}$, in terms of

$$\mathfrak{U} \equiv z \cos \psi, \qquad \mathfrak{V} \equiv z \sin \psi. \tag{183}$$

† Watson[93] p. 379, Equation (1).

Then, the operator $(1 - i\partial/\partial\mathfrak{U})$ applied to $\mathcal{J}$ replaces, in effect, $1 - \cos\Theta$ in the integrand by $\sin^2\Theta$; it gives a Gegenbauer integral† which reduces to $(2\pi/z)^{\frac{1}{2}} \sin\psi\, J_{\frac{3}{2}}(z) = 2\mathfrak{V}z^{-2}(z^{-1}\sin z - \cos z)$. Since this is $-i[\exp(-i\mathfrak{U})]\partial(\mathcal{J}\exp i\mathfrak{U})/\partial\mathfrak{U}$,

$$\mathcal{J} = \mathcal{J}\,|_{\mathfrak{U}=0} \exp(-i\mathfrak{U}) + i2\mathfrak{V}\,[\exp(-i\mathfrak{U})] \cdot \int_0^{\mathfrak{U}} z^{-2}(z^{-1}\sin z - \cos z)\exp(i\mathfrak{U})\,d\mathfrak{U} \tag{184}$$

follows, in which z is $(\mathfrak{U}^2 + \mathfrak{V}^2)^{\frac{1}{2}}$. The integral in this equation may be evaluated by writing $\sin z$ and $\cos z$ in terms of exponentials and introducing $\mathfrak{U} + (\mathfrak{U}^2 + \mathfrak{V}^2)^{\frac{1}{2}}$ and $\mathfrak{U} - (\mathfrak{U}^2 + \mathfrak{V}^2)^{\frac{1}{2}}$ as variables of integration. It is found to equal

$$\tfrac{1}{2}i[\exp(i\mathfrak{U})]\{[\mathfrak{U}^2 + \mathfrak{V}^2 + \mathfrak{U}(\mathfrak{U}^2 + \mathfrak{V}^2)^{\frac{1}{2}}]^{-1}\exp[i(\mathfrak{U}^2 + \mathfrak{V}^2)^{\frac{1}{2}}] + [\mathfrak{U}^2 + \mathfrak{V}^2 - \mathfrak{U}(\mathfrak{U}^2 + \mathfrak{V}^2)^{\frac{1}{2}}]^{-1}\exp[-i(\mathfrak{U}^2 + \mathfrak{V}^2)^{\frac{1}{2}}]\} - i\mathfrak{V}^{-2}\cos\mathfrak{V}.$$

With $\mathcal{J}\,|\,\mathfrak{U}_{=0}$ equal to $2\mathfrak{V}^{-1}(1 - \cos\mathfrak{V})$ from an integral of Sonine,‡ $\mathcal{J}$ is thus found in terms of $\mathfrak{U}$ and $\mathfrak{V}$ and, by use of (180), in terms of κ and U. The respective expressions for F_p of (164) then follow.

REFERENCES

1. Breckenridge, R. G., Russell, B. R. and Hahn, E. E., editors, *Photoconductivity Conference*, John Wiley and Sons, New York, 1956.
2. Hoffmann, A., in Scottky, W., editor, *Halbleiterprobleme*, Vol. 2, Friedr. Vieweg und Sohn, Braunschweig, Germany, 1955, Ch. 5.
3. Fan, H. Y., in Seitz, F. and Turnbull, D., editors, *Solid-State Physics*, Vol. 1, Academic Press, New York, 1955, p. 354.
4. Shulman, R. G., in Hannay, N. B., editor, *Semiconductors*, Reinhold Publishing Corp., New York, 1959, Ch. 11.
5. Bemski, G., Proc. I.R.E., **46,** 1958, p. 990.
6. Bonch-Bruevitch, V. L., Zh. Tekh. Fiz., **28,** 1958, p. 67; translation: Soviet Phys. Tech. Phys., **3,** 1958, p. 60.
7. Sandiford, D. J., Proc. Phys. Soc. (London), **71A,** 1958, p. 1002.
8. van Roosbroeck, W., to be published.
9. van Roosbroeck, W., Bull. Am. Phys. Soc., **2,** 1957, p. 152.
10. van Roosbroeck, W., Phys. Rev., **91,** 1953, p. 282.
11. van Roosbroeck, W., Phys. Rev., **101,** 1956, p. 1713.
12. Schottky, W. and Spenke, E., Wiss. Veröff. aus die Siemens-Werken, **18,** 1939, p. 1.
13. Spenke, E., *Elektronische Halbleiter*, Springer-Verlag, Berlin, 1955, p. 304.
14. Rose, F. W. G., Proc. Phys. Soc. (London), **71B,** 1958, p. 699.
15. Landsberg, P. T., Proc. Phys. Soc. (London), **70B,** 1957, p. 282.
16. Landsberg, P. T., Proc. Phys. Soc. (London), **69B,** 1956, p. 1056.
17. Champness, C. H., Proc. Phys. Soc. (London), **69B,** 1956, p. 1335.
18. Okada, J., J. Phys. Soc. Japan, **12,** 1957, p. 1338.
19. Shockley, W. and Last, J. T., Phys. Rev., **107,** 1957, p. 392.

† Watson[93], p. 379, Equation (1).
‡ Watson[93], p. 374, Equation (3); p. 333, Equation 3.

20. Mercuroff, W., Compt. rend., **246,** 1958, p. 1175.
21. Khartsiev, V. E., Zh. Tekh. Fiz., **28,** 1958, p. 1651; translation: Soviet Phys. Tech. Phys., **3,** 1958, p. 1522.
22. Sah, C.-T. and Shockley, W., Phys. Rev., **109,** 1958, p. 1103.
23. Bernard, M., J. Elect. & Cont., **5,** 1958, p. 15.
24. Kalashnikov, S. G. and Tissen, K. P., unpublished report (S.L.A. translation pool, John Crerar Library, Chicago).
25. Kalashnikov, S. G., J. Phys. Chem. Solids, **8,** 1959, p. 52.
26. Kalashnikov, S. G., Zh. Tekh. Fiz., **26,** 1956, p. 241; translation: Soviet Phys. Tech. Phys., **1,** 1956, p. 237.
27. Mashovets, T. V., Zh. Tekh. Fiz., **28,** 1958, p. 1140; translation: Soviet Phys. Tech. Phys., **3,** 1959, p. 1062.
28. Shockley, W., Proc. I.R.E., **46,** 1958, p. 973.
29. Rose, A., Phys. Rev., **97,** 1955, p. 322.
30. Rose, A., in Gibson, A. F., editor, *Progress in Semiconductors*, Vol. 2, John Wiley and Sons, New York, 1957.
31. Shockley, W. and Read, W. T., Phys. Rev., **87,** 1952, p. 835.
32. Hall, R. N., Phys. Rev., **83,** 1951, p. 228; **87,** 1952, p. 387.
33. Tolpygo, K. B. and Rashba, E. I., Zh. Eksptl. Teoret. Fiz., **31,** 1956, p. 273; translation: Soviet Phys. JETP, **4,** 1957, p. 213.
34. Rose, A., RCA Rev., **12,** 1951, p. 362.
35. Rose, A., Proc. I.R.E., **43,** 1955, p. 1850.
36. Bube, R. H., J. Phys. Chem. Solids, **1,** 1957, p. 234.
37. Fan, H. Y., Phys. Rev., **92,** 1953, p. 1424; **93,** 1954, p. 1434.
38. Adirovich, E. I. and Guro, G. M., Doklady Akad. Nauk SSSR, **108,** 1956, p. 417; translation: Soviet Phys. Doklady, **1,** 1956, p. 306.
39. Sandiford, D. J., Phys. Rev., **105,** 1957, p. 524.
40. Clarke, D. H., J. Elect. & Cont., **3,** 1957, p. 375.
41. Wertheim, G. K., Phys. Rev., **109,** 1958, p. 1086.
42. Isay, W.-H., Ann. Physik, **13,** 1953, p. 327.
43. Nomura, K. C. and Blakemore, J. S., Phys. Rev., **112,** 1958, p. 1607.
44. Guro, G. M., Zh. Eksptl. Teoret. Fiz., **33,** 1957, p. 158; translation: Soviet Phys. JETP, **6,** 1958, p. 123.
45. Okada, J., J. Phys. Soc. Japan, **13,** 1958, p. 793.
46. Blakemore, J. S., Phys. Rev., **110,** 1958, p. 1301.
47. Iglitsyn, M. I., Kontsevoi, Y. A. and Sidorov, A. I., Zh. Tekh. Fiz., **27,** 1957, p. 2461; translation: Soviet Phys. Tech. Phys., **2,** 1958, p. 2292.
48. Burton, J. A., Morton, J. A., Severiens, J. C. and Hull, G. W., J. Phys. Chem., **57,** 1953, p. 853.
49. van Roosbroeck, W. and Shockley, W., Phys. Rev., **94,** 1954, p. 1558.
50. Brill, P. H. and Schwartz, R. F., Phys. Rev., **112,** 1958, p. 330.
51. Stöckmann, F., Z. Phyzik, **143,** 1955, p. 348.
52. Collins, R. J., private communication.
53. Nixon, J. D. and Machlup, S., Bull. Am. Phys. Soc., **3,** 1958, p. 427.
54. Hornbeck, J. A. and Haynes, J. R., Phys. Rev., **97,** 1955, p. 311.
55. Lax, M., J. Phys. Chem. Solids, **8,** 1959, p. 66.
56. Collins, C. B., Carlson, R. G. and Gallagher, C. J., Phys. Rev., **105,** 1957, p. 1168.
57. Bemski, G., Phys. Rev., **111,** 1958, p. 1515.
58. Lashkarev, V. E., Rashba, E. I., Romanov, V. A. and Demidenko, Z. A., Zh. Tekh. Fiz., **28,** 1958, p. 1853; translation: Soviet Phys. Tech. Phys., **3,** 1959, p. 1707.
59. Mironov, A. G., Fiz. Tverdovo Tela, **1,** 1959, p. 525; translation: Soviet Phys. Solid State, **1,** 1959, p. 471.
60. Jonscher, A. K., Proc. Phys. Soc. (London), **70B,** 1957, p. 230.
61. Zitter, R. N., Phys. Rev., **112,** 1958, p. 852.
62. Amith, A., Bull. Am. Phys. Soc., **4,** 1959, p. 28.
63. Amith, A., Phys. Rev., **116,** 1959, p. 793.
64. van Roosbroeck, W., B.S.T.J., **29,** 1950, p. 560.
65. Fan, H. Y., private communication.
66. Jonscher, A. K., Proc. Phys. Soc. (London), **70B,** 1957, p. 223.
67. Kaiser, W., private communication.

68. Rose, F. W. G., Proc. Phys. Soc. (London), **70B,** 1957, p. 801.
69. Landsberg, P. T., Proc. Phys. Soc. (London), **65A,** 1952, p. 604.
70. Ridout, M. S., in *Report of the Meeting on Semiconductors,* Rugby, 1956, The Physical Society, London, 1957, p. 33.
71. Newman, R., Woodbury, H. H. and Tyler, W. W., Phys. Rev., **102,** 1956, p. 613.
72. Morin, F. J. and Maita, J. P., Phys. Rev., **96,** 1954, p. 28.
73. Debye, P. P. and Kohane, T., Phys. Rev., **94,** 1954, p. 724.
74. Buck, T. M., private communication.
75. Haynes, J. R., private communication.
76. Hannay, N. B., Haynes, J. R. and Shulman, R. G., Phys. Rev., **96,** 1954, p. 833.
77. Haynes, J. R. and Hornbeck, J. A., Phys. Rev., **100,** 1955, p. 606.
78. Shulman, R. G., private communication.
79. Kaiser, W., Bull. Am. Phys. Soc., **3,** 1958, p. 409.
80. Kaiser, W., Frisch, H. L. and Reiss, H., Phys. Rev., **112,** 1958, p. 1546.
81. Kaiser, W., Keck, P. H. and Lange, C. F., Phys. Rev., **101,** 1956, p. 1264.
82. Kaiser, W., Phys. Rev., **105,** 1957, p. 1751.
83. Green, G. W., Hogarth, C. A. and Johnson, F. A., J. Elect. & Cont., **3,** 1957, p. 171.
84. Fuller, C. S. and Doleiden, F. H., J. Appl. Phys., **29,** 1958, p. 1264.
85. Fuller, C. S., Ditzenberger, J. A., Hannay, N. B. and Buehler, E., Phys. Rev., **96,** 1954, p. 833.
86. Fuller, C. S., Ditzenberger, J. A., Hannay, N. B. and Buehler, E., Acta Met., **3,** 1955, p. 97.
87. Fuller, C. S. and Logan, R. A., J. Appl. Phys., **28,** 1957, p. 1427.
88. Hannay, N. B., private communication.
89. van Roosbroeck, W. and Buck, T. M., in Biondi, F. J., editor, *Transistor Technology,* Vol. 3, D. Van Nostrand Co., Princeton, N. J., 1958, Ch. 9.
90. Amith, A., private communication.
91. Doetsch, G., *Tabellen zur Laplace-Transformation und Anleitung zum Gebrauch,* Springer-Verlag, Berlin, 1947, p. 57.
92. van der Pol, B. and Bremmer, H., *Operational Calculus Based on the Two-Sided Laplace Integral,* Cambridge Univ. Press, Cambridge, 1955.
93. Watson, G. N., *A Treatise on the Theory of Bessel Functions,* Cambridge Univ. Press, Cambridge, 1952.
94. Pollak, H. O., private communication.

The Charge and Potential Distributions at the Zinc Oxide Electrode

By J. F. DEWALD

(Manuscript received August 20, 1959)

Capacitance measurements made on single crystal zinc oxide electrodes in contact with aqueous electrolytes are reported. Over a wide range of bias and bulk donor density, the results are in almost quantitative accord with predictions of the simple Poisson-Boltzmann (Poisson-Fermi in the degenerate case) equation. This is shown to imply the complete absence of surface-state effects in this system. A very sharp discontinuity in the flat-band potential is observed at bulk electron densities in the range from 0.6 $\times$ 10^{18} to 2 $\times$ 10^{18} cm^{-3}. This and other effects, arising under varying surface treatments, are discussed in some detail. The use of the semiconductor/electrolyte interface in studying the properties of low-lying donors is illustrated for the case of boron, which is shown to lie about 0.3 ev below the conduction band.

I. INTRODUCTION

The modern era in the study of the semiconductor electrolyte interface dates back to Brattain and Garrett's work[1] on the germanium electrode. Practically all of the studies since that time have been on germanium, this work being summarized in two recent review articles.[2,3] Although the qualitative understanding of the germanium electrode is fairly well advanced, complications arise because of the chemical reactivity of the material. This leads, in aqueous solutions at least, to the presence of oxide films and, in addition, precludes study of the electrode under equilibrium conditions.*

Zinc oxide was chosen for the present study because, of all the thermodynamically stable semiconductors (in contact with aqueous solutions), its bulk properties (band structure, mobilities, impurity ionization energies, etc.) are now the most thoroughly understood. We are concerned in the present paper with the distributions of charge and potential at the zinc oxide electrolyte interface.

* See, for example, the work of Bohnenkamp and Engel.[4]

Fig. 1 shows a diagram of a typical n-type semiconductor electrode in contact with an electrolyte under conditions of anodic bias (bands bending up). The conventional energy-band diagram for the system is also shown in Fig. 1. In general, the Galvani (or inner) potential difference, ψ_0, may distribute itself over three regions of the interface: (a) the space-charge layer in the semiconductor, typically 10^{-4} cm or more thick; (b) the one or two atom-diameter region of the Helmholtz layer; (c) the diffuse Gouy layer in the electrolyte. Since the Gouy layer has been much investigated in studies of metal electrodes and is no different at a semiconductor electrode, we will restrict ourselves in both our experimental and theoretical treatments to conditions (namely high ionic concentration in the electrolyte) under which the potential drop across the Gouy layer is negligible. The potential drop across the space-charge layer in the semiconductor, ψ_s, is called the *surface potential.*

The nature of the charge and potential distributions has been determined by measuring the differential capacitance of the zinc oxide electrode as a function of bias. The methods used are identical in principle to those which have been used with metal electrodes to distinguish

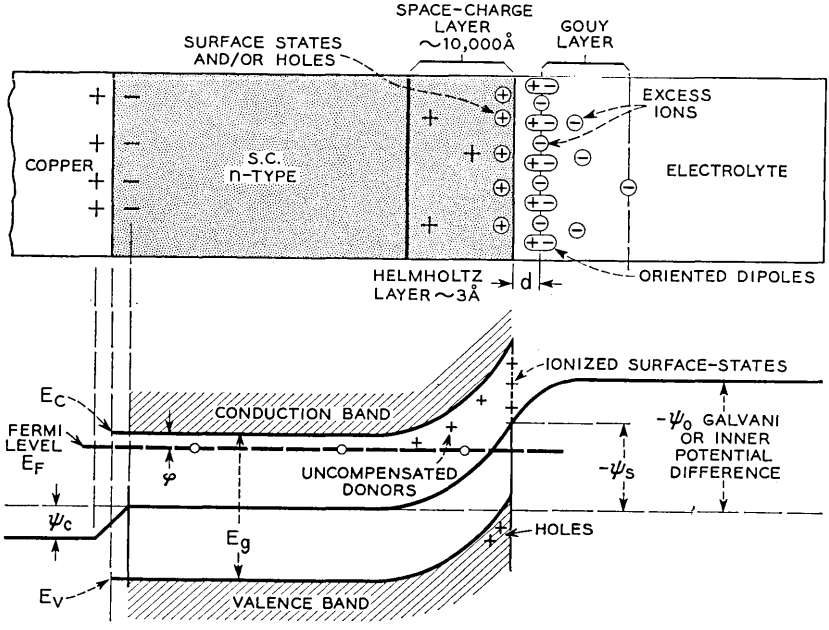


Fig. 1 — Schematic structure and energy-level diagram of a typical n-type semiconductor/electrolyte interface. By convention, one takes the potential in the bulk of the semiconductor to be zero, using the center of the gap as reference point.

"specifically adsorbed" charge residing in the Helmholtz layer from the charge in the Gouy layer.[5] Using the nomenclature of Fig. 1, the capacitance, C, of the electrode may be written*

$$C = -\frac{dQ}{d\psi_0} = -\left(\frac{dQ_{sc}}{d\psi_s} + \frac{dQ_{ss}}{d\psi_s}\right)\frac{d\psi_s}{d\psi_0}, \tag{1}$$

where Q, Q_{sc} and Q_{ss} are, respectively, the total charge on the semiconductor, the charge in the space-charge layer and the charge in "surface states."† For the assumed condition of a negligible potential drop across the Gouy layer, there may still be a sizable variation of the potential drop across the Helmholtz layer if the surface-state density is large. This potential drop may be estimated by taking

$$\frac{d\psi_s}{d\psi_0} = 1 - \frac{C}{C_H}, \tag{2}$$

where C_H is the capacitance of the Helmholtz layer, which should be of the same order of magnitude as that at a metal electrode ($\sim$30 μf/cm^2 on mercury)[5] and thus, anticipating our experimental results, very much larger than C over the entire range of potentials studied. Equations (1) and (2) may thus be combined to yield

$$C = -\left\{\frac{dQ_{sc}}{d\psi_s} + \frac{dQ_{ss}}{d\psi_s}\right\}\left(1 - \frac{C}{C_H}\right) \approx -\left(\frac{dQ_{sc}}{d\psi_s} + \frac{dQ_{ss}}{d\psi_s}\right). \tag{3}$$

This may be written in the form

$$\frac{1}{C} = \frac{1}{C_H} + \frac{1}{C_{sc} + C_{ss}}, \tag{4}$$

where C_{sc} and C_{ss} — in effect, capacitances of the space-charge layer and of the surface states — are defined respectively as $dQ_{sc}/d\psi_s$ and $dQ_{ss}/d\psi_s$.

* Unfortunately, the surface physicist's convention for the sign of potentials is not the same as that of the electrochemist. The physicist says that the more the bands bend up, i.e., the more anodic the bias, the more *negative* is the surface potential. The electrochemist says that the more anodic the bias, the more *positive* the electrode potential. We shall use *both* conventions, according to the context of the statement involved, using unambiguous phrases such as "anodic," "cathodic," "bands bent up," "bands bent down" when confusion is particularly probable.

† Surface-states on a semiconductor were first invoked by Bardeen[6] in an attempt to understand the fact, determined from field-effect conductivity measurements, that the electric field did not penetrate as far into the semiconductor as simple theory said it should. In effect, surface states are analogous to "specifically adsorbed" charge at a metal electrode. The major difference is that surface states equilibrate with the *electrode* rather than with the electrolyte. Law[7] gives a review of recent work on surface states. For a given crystal and surface condition, Q_{ss} is assumed to be uniquely determined by ψ_s .

With these definitions, (4) shows that the total capacitance may be represented by an equivalent circuit with a capacitance C_H in series and two capacitances, C_{sc} and C_{ss}, in parallel. Now C_{sc} may be computed from first principles (see below) and C_H estimated from results with metal electrodes, so that a measurement of C gives an estimate of C_{ss}.

II. THEORY OF THE SPACE-CHARGE CAPACITANCE FOR A SINGLE IMMOBILE DONOR

The theory of the space-charge layer at a semiconductor surface has been given by a number of workers.[8,9,10] We derive here the space-charge capacitance for the case of a large-energy-gap n-type semiconductor with immobile donors that may, however, be only partially ionized. A number of our crystals contained low-lying donors, and the results are understandable only in terms of this more elaborate treatment, which includes strongly dissociated donors as a special case. The Poisson-Boltzmann equation is used, consideration of the degenerate surface being given in a separate treatment below. The situation here is quite analogous to that existing in the Gouy layer with the one exception that one of the charged species is immobile.*

The Poisson-Boltzmann equation for the system may be written

$$\frac{d^2y}{dx^2} = \frac{q^2}{\kappa\epsilon_0 kT}\,(n_0 e^y - N^+), \tag{5}$$

where n_0 is the electron density in the bulk of the semiconductor, N^+ is the density of *ionized* donors at the point x (measured from the surface) and y is the potential, measured from the bulk, in units of kT/q. This assumes that the electrons are in equilibrium all across the space-charge layer and that, because of the large energy gap, the space charge due to holes is negligible. We also assume that at each point the donor ionization reaction

$$M \leftrightharpoons M^+ + e^- \tag{6}$$

is in equilibrium with the *local* electron density (ne_0^y). The equilibrium constant, K, for this reaction may be written in the form[12]

$$\frac{nN^+}{N_D - N^+} = N_c\left[\frac{m^{(N)}}{m}\right]^{\frac{3}{2}} D^{-1}\exp\left(\frac{-E_D}{kT}\right) = K, \tag{7}$$

where N_D is the *total* density of donors, N_c is the free electron density-of-states $[2(2mkT/h^2)^{\frac{3}{2}}]$, m is the electron mass, $m^{(N)}$ is the density-of-states effective mass, D is the donor degeneracy, and E_D is the energy

* Macdonald[11] has treated this problem previously. [Note added in proof.]

difference between the donor level and the bottom of the conduction band.

K and N_D are constant across the space-charge layer. However, because of the electric field, the electron density (n) and the *ionized* donor density (N^+) vary in a manner consistent with (7). Inserting (7) into (5), integrating once with respect to y, and using the boundary condition that in the bulk of the crystal y and dy/dx are both zero, we obtain the electric field at any point, x, in the form

$$\frac{dy}{dx} = \pm \left(\frac{2q^2N_D}{\kappa\epsilon_0 kT}\right)^{\frac{1}{2}} \{f(e^y - 1) - y + \ln[f + (1 - f)e^y]\}^{\frac{1}{2}}, \quad (8)$$

where

$$f = \frac{N_0^+}{N_D};$$

i.e., f is the fraction of the donors that are ionized in the bulk of the semiconductor. The positive sign in (8) applies when the bands bend up, the negative sign when they bend down.

Now the space-charge capacitance is the derivative of the space charge, Q_{sc}, with respect to the surface potential, with Q_{sc} in turn being given by

$$Q_{sc} = \kappa\epsilon_0 \frac{kT}{q} \left(\frac{dy}{dx}\right)_{x=0}. \quad (9)$$

The space-charge capacitance is obtained by substituting (8) into (9) and differentiating:

$$C_{sc} = -\frac{dQ_{sc}}{d\psi_s} = \left(\frac{q^2N_D\kappa\epsilon_0}{2kT}\right)^{\frac{1}{2}} \frac{\left| fe^Y - \dfrac{1}{1 + \left(\dfrac{1}{f} - 1\right)e^Y} \right|}{\{f(e^Y - 1) - Y + \ln[f + (1 - f)e^Y]\}^{\frac{1}{2}}}, \quad (10)$$

where Y is the value of the surface potential in units of kT/q, i.e., the value of y at $x = 0$. If the donors are completely ionized in the bulk of the semiconductor, $f = 1$ and (10) reduces to

$$C_{sc} = \left(\frac{q^2N_D\kappa\epsilon_0}{2kT}\right)^{\frac{1}{2}} \frac{|e^Y - 1|}{\{e^Y - Y - 1\}^{\frac{1}{2}}}. \quad (11)$$

A number of special cases of (10) are of specific interest. When Y is strongly negative; i.e., when the bands are bent up quite strongly, we have

$$C_{sc} = \left(\frac{q^2N_D\kappa\epsilon_0}{2kT}\right)^{\frac{1}{2}} \{-f + \ln f - Y\}^{-\frac{1}{2}}. \quad (12)$$

In the limit of large negative Y this is just the Mott-Schottky approximation.[8,9] Independent of the value of f, a plot of $1/C_{sc}^2$ versus Y should give a straight line, at sufficiently negative Y, whose slope is a measure of the *total* donor density at the surface. The only implied requirement is that the electrons be in equilibrium across the space-charge layer, i.e., that the Fermi level be flat right up to the surface. The maximum value by which the bands can be bent up while still keeping the Fermi level flat depends upon the magnitude of the current flow across the semiconductor/electrolyte interface. While this may be very small (less than 10^{-10} amperes/cm^2 in typical experiments on zinc oxide), it is still finite, and in consequence the Fermi level, under increasing anodic bias, eventually starts to follow the bands. At best, (12) is expected to apply for donors located less than one volt or so below the Fermi level.

At the other extreme of potential, when the bands bend down and Y is large and positive, (10) may be approximated in the form

$$C_{sc} = \left(\frac{q^2 N_D K \epsilon_0}{2kT}\right)^{\frac{1}{2}} (fe^Y)^{\frac{1}{2}}. \tag{13}$$

As can be seen, the capacitance increases, apparently without limit, as the bands bend down, and the electrons tend to crowd up closer and closer to the surface. Eventually, however, the surface layers become degenerate and then the Poisson-Boltzmann approximation breaks down. This region is considered in greater detail below. The behavior of (10) in the region of intermediate potential is fairly complex, particularly for small values of f, and is also considered below.

III. EXPERIMENTAL PROCEDURES

3.1 *Sample Preparation*

The crystals used in this study were grown in these laboratories by D. G. Thomas and R. T. Lynch using the vapor-phase reaction between zinc and oxygen at about 1200°C, as originally reported by Scharowsky.[13] The resulting crystals were in the form of hexagonal needles, about 0.1 to 0.3 mm in "diameter" with the primary faces being {11.0}. Conductivities of the "as-grown" crystals were in the range from about 0.01 to 3.0 ohm^{-1} cm^{-1}. The crystals were selected for uniformity of composition, both lengthwise and radial. The lengthwise uniformity was checked by conductivity measurement, while the radial uniformity was checked by measurement of both conductivity and electrode capacitance between successive etches in 85 per cent H_3PO_4. Of a number of etching solu-

tions employed* this was found to be the most satisfactory by virtue of its uniform attack and its convenient rate (~1 micron per minute at room temperature).

In the lowest region of conductivity the crystals were used in the as-grown condition. For conductivities in excess of about 1 ohm^{-1} cm^{-1}, the crystals were "doped" by high-temperature diffusion of either hydrogen or indium into the crystal. Although it is the more convenient donor because of its higher diffusion coefficient,[14] hydrogen was not used extensively, since it was found to be too mobile, even at room temperature, under the very high electric fields near the surface.

Most of the crystals were doped with indium in a manner similar to that described by Thomas.[15]† For conductivities in excess of about 15 ohm^{-1} cm^{-1}, the crystals were equilibrated with an excess of $In(NO_3)_3$ (applied by dipping the crystal into an aqueous solution) at various temperatures in the range from 950° to 1150°C for periods up to about 20 days.‡ The crystals become saturated with respect to indium, the concentration depending upon the particular temperature selected. This procedure was rather tedious for conductivities below about 10 ohm^{-1} cm^{-1} because of the extended equilibration times that would have been required by virtue of the low temperatures. In this conductivity range the crystals were partially equilibrated with an excess of $In(NO_3)_3$ until the average conductivity had the desired value. The excess indium was then removed by etching in H_3PO_4, and the crystals were replaced in the oven, usually at 1100°C, for times calculated (from Thomas' diffusion data[15]) to give a maximum radial variation of no more than 20 per cent in donor density.

For some reason (not understood at present) a sizable fraction of the crystals treated in this manner showed very large radial nonuniformity, variations in conductivity as large as a factor of ten being not uncommon. Extending the equilibration time by factors of ten or more did not improve matters, the average conductivity increasing while the nonuniformity remained large. Other crystals, however, behaved quite nicely in accord with Thomas' diffusion data and, except for a surface skin about

---

* Other etches employed included HF, HNO_3 (both dilute and concentrated), HCl and KCN. The KCN etch is unique in developing quite nice {21.0} crystal faces even though one starts with {11.0} faces initially; H_3PO_4 preserves the original orientation.

† The indium ions enter the crystal substitutionally with an activation-energy-for-diffusion of about 3 ev compared to an activation energy of about 0.9 ev for the interstitial hydrogen.[14,15] Consequently, they are much less mobile.

‡ The oven employed was platinum-wound and had a recrystallized alumina lining to try to minimize contamination. The crystals were supported on other crystals of zinc oxide contained in "boats" of recrystallized alumina.

a micron thick, could be readily equilibrated. The only explanation of these effects that seems at all possible is that the indium diffusion is structure-sensitive (e.g., due to a varying dislocation density) and that a second donor (perhaps oxygen vacancies) is introduced when one equilibrates for times much in excess of the values calculated from Thomas' data. Whatever the cause of this effect, only those crystals were used that showed less than a 50 per cent variation in donor density over the entire crystal. The electron and donor densities reported below are appropriate averages.

Mechanically strong ohmic contacts were made to the crystals by electroplating a thin layer of indium on the tip of the crystal, then electroplating a second layer of copper on top of the indium and, finally, soldering a copper wire to the copper plate. A fairly large "glob" of solder was used, which completely surrounded the tip of the crystal. This clamps the crystal so tightly that, on an attempt to break the contact, the crystal breaks instead. The resistance of the contact was negligible compared to the crystal resistance over the entire range of conductivity from about 0.005 to 150 ohm^{-1} cm^{-1}.*

The solder joint and copper wire were masked by slipping the copper wire through a pyrex glass tube and then sealing with Apiezon W wax. The efficacy of the masking procedure was checked by measuring the direct current flow under anodic bias up to about 10 volts. If any of the copper, solder or indium is exposed to the solution, catastrophic currents flow under such a bias. On the other hand, a properly masked crystal passes anodic currents only of the order of 10^{-9} amperes per cm^2 or less, almost completely independent of the crystal resistivity.†

3.2 *Electrochemical Cell and Capacitance Bridge*

The electrochemical cell employed was of fairly standard design. It was made of quartz and contained a large-area working electrode of platinized platinum, a normal calomel electrode separated from the main solution by a 1 normal KCl salt bridge, and a second platinum electrode for preelectrolysis. A nitrogen atmosphere was maintained over the solution at all times. The solution was a borate-buffered 1 normal solution of recrystallized KCl (pH $\sim$ 8.5). Freshly ground, spectroscopically pure

* A more convenient but less satisfactory technique for making contact was to fuse a platinum wire directly to the crystal. This gives a nice low-resistance contact of good mechanical strength. However, the high temperature involved introduces impurities of an unknown nature and many of the results with such contacts were anomalous, particularly on crystals that were heat-treated subsequent to the platinum fusion.

† A study of the electrode kinetics at the zinc oxide electrode has been reported on[16] and details will be published shortly.

zinc oxide with an estimated area of about 10,000 cm^2 was added to the solution as a "getter" for any surface-active impurities. Periodic additions of more zinc oxide were made.

Impedance measurements were made using a General Radio #716-C bridge with an auxiliary decade-capacitance box in parallel with the variable air capacitor of the bridge. A schematic diagram of the circuit is shown in Fig. 2. The capacitance of the zinc oxide electrode, C_x, was always less than about 0.3 μf. The platinum working electrode had an apparent area of about 10 cm^2, so that its capacitance was more than 10^3 times C_x and thus made a negligible contribution to the total capacitance.

The equivalent circuit for the zinc oxide electrode is shown schematically in Fig. 2 as a resistance and capacitance in series. This is appropriate for the highest conductivity crystals studied, where the resistance is small and arises primarily in the electrolyte. For lower conductivity

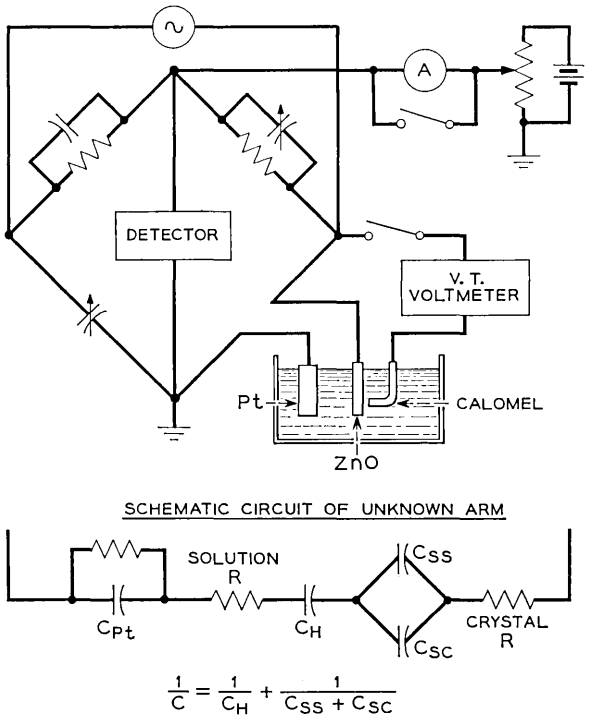


Fig. 2 — Schematic diagram of the capacitance bridge and the equivalent circuit of the unknown arm. The entire immersed area is in contact with the solution, Fig. 1 representing each *element* of area.

crystals the crystal resistance becomes predominant and sizable. Because of the rod-like habit of the crystals, the equivalent circuit is then analogous to a transmission line, the portions of the crystal nearest the tip having a larger resistance in series with them than do the regions near the metal contact. In most of the data presented below, the frequency was kept low enough to make the transmission-line correction negligible. Corrections, never over about 10 per cent, were made in the study of frequency effects.

The electrode bias applied to the crystal was varied potentiometrically and measured, with respect to the calomel electrode, by a Leeds and Northrup vacuum tube voltmeter with an impedance of 1600 megohms and a sensitivity of about ± 1 mv. The voltmeter was disconnected during capacitance measurement. The input signal to the bridge was kept below 10 mv. Measurements were made in the frequency range from 50 cps to 100 kc. The bridge was extensively calibrated at the start of the experiments, and then checked periodically during the course of the measurements. The major difficulty encountered was in the minimization of the lead capacitance, which was finally reduced to about 12 $\mu\mu$f. Even this small value was significant in the measurements of the lowest conductivity crystals, since the total capacitance in these crystals was as low as 150 $\mu\mu$f.

IV. EXPERIMENTAL RESULTS

4.1 *Exhaustion Region*

The results that have been obtained for the capacitance of the zinc oxide electrode in the exhaustion region, i.e., under anodic bias, are remarkable for their simplicity. The results for two crystals of intermediate conductivity are shown in Fig. 3. Here, taking our cue from (12), we plot the reciprocal of the capacitance squared against the electrode potential. As can be seen, the resulting plots are very nicely linear for the two crystals. This was the case for nearly every crystal studied; in only two crystals out of more than 50 was a curvature greater than 2 per cent observed in the region of positive potential (on the calomel scale).

The precision of the linearity on most crystals is quite remarkable and is illustrated in Table I for one of the crystals shown in Fig. 3. As can be seen, the slope is constant within about 0.5 per cent, the departures being well within the experimental uncertainty. This behavior is precisely that predicted for the Mott-Schottky space-charge capacitance [see (10)] and implies that the surface-state capacitance is very much smaller than the space-charge capacitance. Note in this connection that a capacitance versus bias curve for any *one* crystal can *always* be fitted

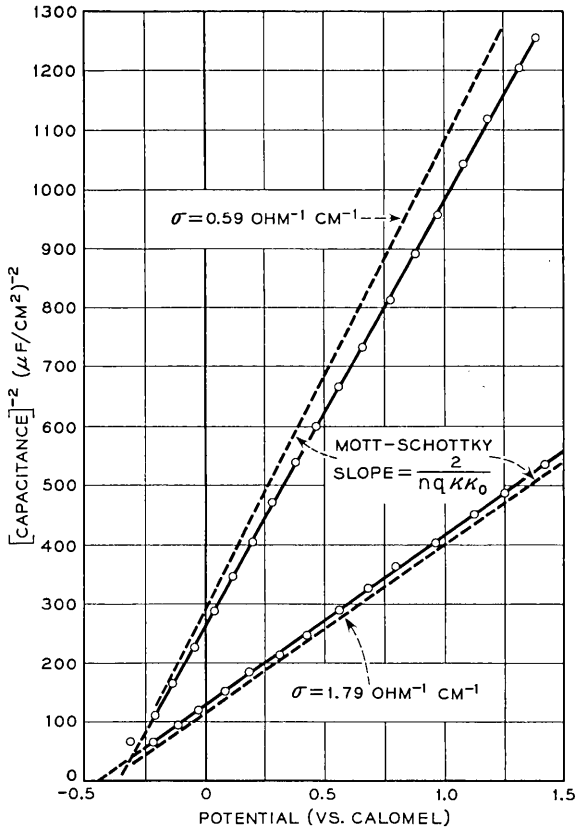


Fig. 3 — Mott-Schottky plots for two crystals under exhaustion conditions; the dotted lines represent the theoretical slopes. At potentials more negative than about −0.3 volt, the capacitance increases exponentially (see Fig. 8). The intercepts of the linear plots afford a quantitative measure of the electrode potential at the flat-band condition.

by a completely unrestricted distribution of surface states, such an interpretation violating only the law of simplicity. However, one cannot devise a single surface-state distribution that will simultaneously satisfy the results on crystals of varying impurity density.

The magnitude of the slope of the $1/C^2$ versus V plots confirms the simplicity of the results in the exhaustion region. Equation (12) says that, in the absence of surface states and if the donors are completely ionized (i.e., $N_D = n$), the slope should be given by

$$\text{slope} = \frac{2}{nq\kappa\epsilon_0}. \tag{14}$$

TABLE I — THE CAPACITANCE OF A TYPICAL CRYSTAL AS A FUNCTION OF BIAS

$\sigma = 0.59$ ohm^{-1}cm^{-1}

V_{cal}	Capacitance, C, in μf/cm^2	$1/C^2$	$\Delta\ 1/C^2$
1.400	0.02890	1197.3	
			253.5
1.000	0.03255	943.8	
			250.6
0.600	0.03798	693.2	
			252.4
0.200	0.04763	440.8	
			252.1
−0.200	0.07280	188.7	

Since the dielectric constant of zinc oxide is known[12] ($\kappa = 8.5$), the only unknown quantity is the electron density n, which can be obtained from conductivity and Hall-effect measurements. For the two crystals in Fig. 3 the agreement between the Mott-Schottky slope and the experimental slope is within experimental error,* as shown by the dotted lines whose slopes are calculated from (14).

In very good approximation the data are frequency independent, as shown for a typical case in Table II.

The conclusion that the surface-state capacitance is negligible compared to the space-charge capacitance is confirmed in all of our experiments. This is shown in Fig. 4, where we plot the slope of the $1/C^2$ versus V curves as a function of electron density. The open circles represent "as-grown" crystals, while the solid circles are for indium-doped samples. The straight line in Fig. 4 is an absolute prediction with no adjustable parameters. As can be seen, the agreement of the data with

TABLE II — THE FREQUENCY DEPENDENCE OF CAPACITY, CRYSTAL C-3

(Corrected for Transmission-Line Effect)

Frequency	Slope of $1/C^2$ vs. V Plot (μf^{-2}cm$^4V^{-1}$)	Flat-Band Potential*
100	626	−0.471
200	626	−0.471
1,000	628	−0.470
5,000	625	−0.483
10,000	629	−0.441

* See below.

* The major errors arise in the determination of surface area and in the determination that the crystals are in fact uniformly doped.

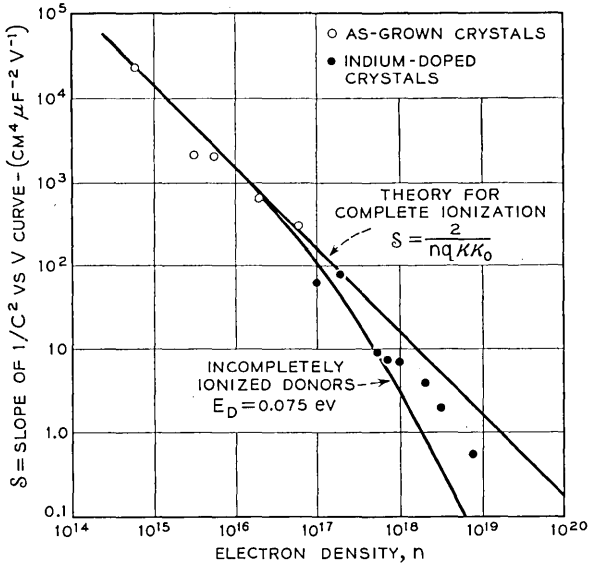


Fig. 4 — The Mott-Schottky slope as a function of the bulk electron density; the straight line is taken from (14) and represents a "no-parameter" fit to the data, the curved line shows the correction for incomplete donor dissociation with E_D = 0.075 ev.

this simple theory is fairly good over the entire range studied. It can be improved somewhat by realizing that at high density the indium donors are not completely ionized. Hutson's preliminary measurements[17] on indium donors indicate an ionization energy of about 0.075 ev at electron densities around 10^{17} or less. If we assume that this value applies over the entire range of concentration and use (7) with[12] $D = 2$ and $m^{(N)}/m = 0.5$, we obtain the curve marked "$E_D = 0.075$ ev." The data lie reasonably well between these extreme predictions.*

The experiments reported above show that the surface-state density is very low in the energy region up to about one volt below the conduction band. We may compute an order-of-magnitude upper limit for their number by considering the data obtained on the lowest conductivity crystal studied. The capacitance versus bias curve for this crystal agreed with that predicted for the simple space-charge layer capacitance within about 2 per cent over the entire exhaustion region. Thus, the total varia-

* The data from a number of crystals have been excluded from Fig. 4 owing to nonuniformity of the donor density. A few others which gave results widely discordant with those of Fig. 4 have also been excluded on the basis of a complex behavior at potentials in the range from 0 to −0.4 volts on the calomel scale. These effects are treated below and are shown to arise from the inadvertent introduction of boron which acts as a low-lying donor.

tion in the charge of the surface-states must have been less than about 2 per cent of the total space charge. The donor density in this crystal was about 0.7×10^{15}, so that the total charge in the space-charge layer [see (10)] at a surface potential of -0.7 volt was 7.7×10^{-9} coulombs/cm^2.* Two per cent of this corresponds to an upper limit for the surface-state density of the order of 10^{+9} states per cm^2, in the energy range from about 0.25 ev to 0.95 volt below the conduction band. (The Fermi level in this crystal lay about 0.25 ev below the conduction band.)

The effective absence of surface states at the zinc oxide/electrolyte interface implies that the electric field arising from free charge on the semiconductor cannot give rise to any significant potential drop across the Helmholtz layer. However, a variable potential drop across this layer can arise if desorbable polar molecules are present at the surface. A typical determination of a $1/C^2$ versus V plot normally took about 10 minutes. The nice linearity of these plots thus precludes any rapid changes with bias of the chemistry (and therefore the surface dipole) of the Helmholtz layer. Sizable slow changes of the surface dipole, on etching a crystal or on long-term equilibration, are indicated, however, by our measurements of the "flat-band potential," i.e., the electrode potential at which the bands are flat. At this potential the charge within the semiconductor is equal to zero.

The flat-band potential is a particularly important quantity with semiconductor electrodes. It is almost perfectly analogous to the potential of the electrocapillary maximum at metal electrodes. Two factors determine its magnitude (for a given reference electrode): the surface dipole at the electrolyte contact and the "built-in" potential drop at the ohmic contact to the lead wire. Both of these are subject to variation, the first by variation of the surface pretreatment and/or the nature of the solution, and the second by variation of the semiconductor doping. Since it is possible to make these variations independently, one can obtain information about both the surface dipole and the bulk Fermi level.

There are several methods of determining the flat-band potential. Surface-state effects having been shown to be absent, one can simply use (10) (in the limit as Y goes to zero) to compute the capacitance at flat band and then adjust the bias until this capacitance is attained. This method has the disadvantage that one must know the surface roughness of the crystal to go from capacitance per unit area, C_{sc}, to the total capacitance.

---

* The value of -0.7 volt is used here because, when the bands bend up much more than this, the Fermi level no longer remains flat and the surface states are no longer in equilibrium with the bulk Fermi level.

An alternative is to use (12). From this, we see that the extrapolation of the linear portion of a $1/C^2$ versus V plot to the voltage axis gives the potential at which the bands bend up by $(f - \ln f)(kT/q)$ volts. A knowledge of f then allows a computation of the flat-band potential, independent of any statements regarding surface roughness.

Fig. 5 shows the experimental variation of the flat-band potential (as determined by the second method described above), with the bulk electron density for three different surface treatments. The curves marked "H_3PO_4" and "KOH" were taken within a few minutes after a brief (2 to 10 second) etch-in for H_3PO_4 (85 per cent) and KOH (3M). The curve marked "long stand" was obtained with crystals that had stood in the buffered KCl electrolyte (pH $\sim$ 8.5) for 10 hours or more.

Consider the data at low concentration first. There are two possible sources of a variable flat-band potential: (a) a variable surface dipole and (b) a variable contact-potential difference at the copper/zinc oxide interface due to a variation in the bulk Fermi level. The latter variation goes as $(kT/q)\ln n$ for a nondegenerate semiconductor, and the straight lines at low electron density are drawn with slope to correspond to this variation (59 mv per decade in n). As can be seen, the change in the "built-in" potential at the copper/zinc oxide interface accounts for most,

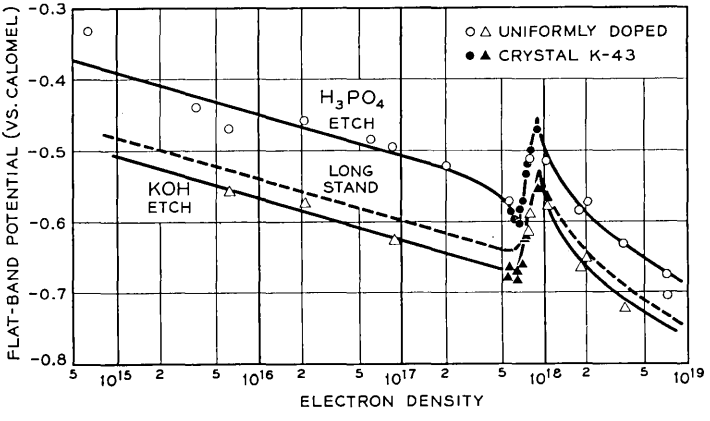


Fig. 5 — The variation of the flat-band potential with bulk electron density for three different surface treatments. The lines at low density are drawn with slope of 59 mv per decade, corresponding to the classical variation of the Fermi level with electron density. Each unshaded point represents a different crystal and/or a different surface treatment. The shaded points are the data obtained by successive etching of a crystal (K-43) containing a (purposely introduced) radial gradient of concentration. For these points the concentration is that just to the semiconductor side of the space-charge layer. The anomalous "bump" in the curves is thought to represent a similar discontinuity in Fermi level.

if not all, of the observed variation with crystal doping at levels below about 0.6×10^{18} electrons per cm^3. This is not particularly surprising, since the impurity density at such levels is quite low in normal chemical parlance (99.99 mole per cent ZnO) and one would not expect the chemistry of the surface to be strongly impurity-sensitive at these low levels.

A more physically interesting result is obtained from the variation of the flat-band potential with the nature of the surface treatment. As can be seen from Fig. 5, the flat-band potential becomes more negative by about 130 mv when one etches in KOH instead of H_3PO_4, corresponding to the addition of a surface dipole with the positive end towards the electrolyte (or removal of a dipole of opposite polarity). This effect is quite reproducible, on a given crystal, and quite rapid; a two-second etch in either H_3PO_4 or KOH (corresponding to the removal of about 300 Å and 30 Å in the two cases) brings the flat-band potential to the values shown, completely independent of the past surface treatment. Several crystals were cycled between acid and base as many as eight times, and the maximum deviation of the flat-band potential (for a given treatment) was about 15 mv. Other acids, HCl and HF for example, give the same value of the flat-band potential as H_3PO_4; however, they also roughen the surface appreciably and have therefore not been studied extensively.

The individual data points shown in Fig. 5 were obtained on freshly etched surfaces. If a crystal is allowed to stand in contact with the electrolyte, a slow change in flat-band potential is observed for upwards of five hours or more. The direction of the change depends upon the nature of the last previous etch, while the final limiting value of the flat-band potential is independent of the nature of the etch. Fig. 6 shows the typical transient behavior following a series of etches in H_3PO_4. Time

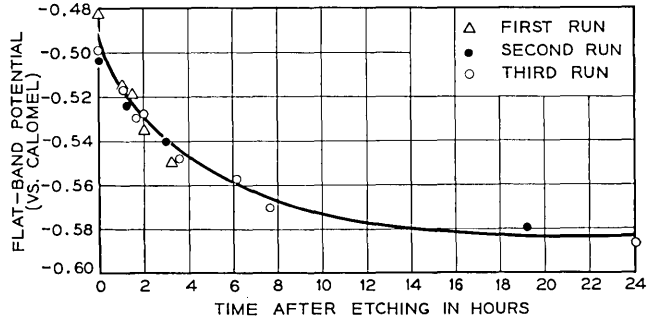


Fig. 6 — The variation of the flat-band potential with time after etching in H_3PO_4; various points correspond to three successive runs on same crystal.

constants for the effect are typically of the order of three hours, independent of crystal doping and also of the bias under which the crystal stands.

The 130-mv change in flat-band potential observed on going from a KOH-etched surface to an H_3PO_4-etched surface may be qualitatively understood as arising from a heterogeneous acid-base equilibrium. On an ideal {11.0} surface there are equal numbers of zinc and oxygen atoms in the surface plane. In the covalent model of the "clean" crystal surface each of these atoms has a "dangling" bond, while in the ionic structure each surface oxygen has an unshared pair of electrons pointing into the solution and each zinc has an empty pair of orbitals. The true situation is probably intermediate between these extremes, but in either case the oxygen will be able to accept a proton, becoming an OH^- ion, and the zinc will be able to complete its tetrahedral coordination by bonding with any ion or molecule from the solution that has an unshared pair of electrons.

A diagram of the surface structure expected when the {11.0} crystal face is in contact with a solution of HCl is shown in Fig. 7. Here we have drawn the zinc, oxygen and chloride atoms as if they were spheres of radius equal to Pauling's tetrahedral covalent radii, and the hydrogen is drawn as a sphere of radius sufficient to make the O–H distance equal to the bond length in the hydroxide ion (~1.0 Å). As can be seen, the proton gets much closer to the crystal than does the chloride ion. To the extent that the bonds have the same degree of ionic character — a not unreasonable approximation in view of the respective electronega-

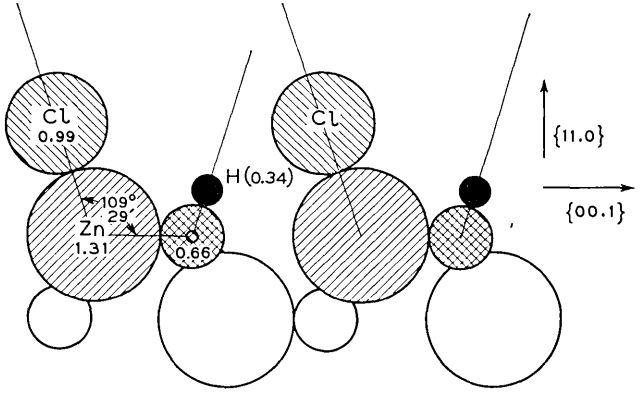


Fig. 7 — Diagram of the surface configuration on a {11.0} surface of ZnO in contact with a solution of HCl; the various distances are computed from Pauling's tetrahedral covalent radii.

tivities — this would then give a surface dipole with the negative side towards the solution. If such a surface is now treated with a strong base like KOH, the protons will tend to be removed, to be replaced by either water molecules or, conceivably, potassium ions. In either case, the surface dipole should become more negative. This is the direction of the change in flat-band potential observed on going from an acid to a basic etch. At intermediate values of pH we expect intermediate values of the flat-band potential corresponding to partial coverage of the surface with protons. This is qualitatively in accord with the data in Fig. 5, the flat-band potential at pH 8.5 being intermediate between the acid and basic treatments.

The one difficulty here is in understanding the very slow equilibration rate. The exchange of protons and chloride ions should presumably be quite rapid, and we cannot think of any other rate-limiting process. The large number of presently unverifiable assumptions involved (about electronegativity, dielectric saturation, etc.) makes it unprofitable to attempt a quantitative calculation of the size of the acid-base effect.

Consider now the data in Fig. 5 at high doping levels. A remarkably sharp "bump," amounting almost to a discontinuity, in the flat-band potential is observed centered at a bulk electron density of about 10^{18}, with the flat-band potential getting more *positive* by about 130 mv on going from an electron density of 0.6×10^{18} to 0.9×10^{18}. Within experimental error the contour of the bump is independent of surface treatment. The "width" of the bump is so small that in our early experiments, which had only one crystal in the range from 0.6×10^{18} to 3×10^{18}, we completely overlooked it and took the low-density behavior to be valid over the entire concentration range. We had then just a single anomalous point, which we rationalized as arising from some unknown anomaly in chemical or thermal history of the crystal. This rationalization became quite tenuous as further crystals were studied in this concentration range, and it was completely eliminated by the data shown as shaded points in Fig. 5. To obtain these data we took a crystal (K-43) that, from capacitance measurements, had a total donor density less than 10^{16} donors per cc and, by a relatively short heat treatment with indium, produced a nonuniform donor distribution. The outside of the crystal had a donor density near 10^{18}, while the inside was at a level around 10^{17} donors per cc. Capacitance and conductivity measurements were then made as the crystal was successively etched. In this way we obtained a plot of flat-band potential versus surface donor density with a minimum of ambiguity as regards uncontrolled variables of thermal and chemical nature. The results are shown in Table III and illustrated graphically in Fig. 5.

TABLE III — THE VARIATION OF THE FLAT-BAND POTENTIAL NEAR THE ANOMALOUS "BUMP", CRYSTAL K-43

Etch Time (minutes in H_3PO_4)	Crystal "Diameter" $= \left(\frac{4}{3} A_{cross}\right)^{\frac{1}{2}}$	Slope of $1/C^2$ vs. V plot*	Intercept in H_3PO_4	Intercept in KOH	Electron Density†
1	0.0238	8.45	−0.441	−0.524	8.7×10^{17}
2	—	9.81	−0.473	−0.559	7.8×10^{17}
3	—	10.30	−0.490	−0.588	7.50×10^{17}
4	—	10.60	−0.504	−0.593	7.25×10^{17}
6	—	11.16	−0.544	−0.630	6.90×10^{17}
8	—	11.89	−0.553	−0.632	6.60×10^{17}
10	—	12.45	−0.574	−0.639	6.40×10^{17}
15	—	12.86	−0.570	−0.656	6.20×10^{17}
25	0.0208	14.05	−0.556	−0.650	5.75×10^{17}
35	—	13.95	−0.563	−0.642	5.80×10^{17}

* On H_3PO_4-etched crystal.
† Electron densities computed from slope of $1/C^2$ vs. V plots using Fig. 4.

As can be seen, these data are nicely consistent with the data obtained on uniformly doped crystals and demonstrate, to our satisfaction at least, that the anomalous variation of flat-band potential with doping is a real effect.

We have no explanation of the anomaly at the present time. The value of the electron density at which it occurs would seem to be unique in one way, 10^{18} electrons per cc marking the onset of electronic degeneracy in zinc oxide. However, it is hard to see how this would play any role in determining the surface dipole. Another possible explanation is that the Fermi level is exhibiting an anomolous variation with impurity density. This would show up as an anomalous variation in the built-in potential difference at the zinc oxide/copper interface.

A few experiments have been made to see the effects of known impurities in the electrolyte upon the capacitance and flat-band potential. The addition of octyl alcohol, which is known to be strongly absorbed at a mercury electrode,[18] has no measurable effect on the capacitance over the entire range of potentials studied.* Nitrobenzene, a strongly polar molecule, was also found to have no effect on the capacitance. Further

* This is perhaps not too surprising on the anodic side of the flat-band potential since in Grahame's study[18] the *potential* of the electrocapillary maximum was not significantly affected by octyl alcohol but only the capacitance, which decreased from about 30 μf per cm^2 to about 4 μf per cm^2. It *is* surprising that no effect was observed at cathodic potentials (see Fig. 8). A comparable change in the capacitance of the Helmholtz layer at the zinc oxide electrode would have been readily detectable at bias 0.3 volt or more cathodic than the flat-band potential where the capacitance is of the order of 5 μf per cm^2.

study of these effects is surely warranted. For the present, we conclude that, although the free charge on the zinc oxide electrode is located almost entirely *inside* the semiconductor, significant portions of observed changes in electrode potential may still occur across the Helmholtz layer.

4.2 *Enrichment Region*

Consider now the data obtained in the enrichment region, i.e., when the bands are bent down. Results typical of the behavior of crystals with completely (or nearly completely) dissociated donors are shown in Fig. 8. As seen from (13), the simple Poisson-Boltzmann theory predicts an exponential dependence of capacitance on bias starting near the flat-band potential. For crystals of conductivity below about 1 ohm^{-1} cm^{-1},

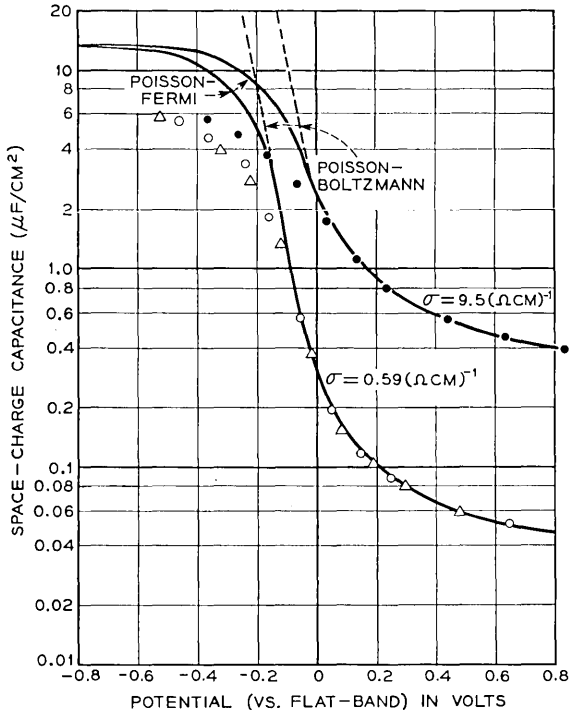


Fig. 8 — The capacitance versus bias curves for two typical crystals at potentials near and cathodic to the flat-band potential; the theoretical curves marked Poisson-Boltzmann and Poisson-Fermi involve only one unknown parameter, the flat-band potential, with is determined from plots like Fig. 3. At points indicated by triangles electrolyte contained a few drops of octyl alcohol.

such an increase is in fact observed in the potential region between where the bands are flat and where the conduction band at the surface goes degenerate.* However, with crystals of conductivity much higher than 1 ohm^{-1} cm^{-1} this region is so narrow as to make the limiting form of (13) invalid.

The curves labeled "Poisson-Boltzmann" in Fig. 8 were calculated from equation (10) and contain no new adjustable parameters. Sizable departure of the experimental data from the Poisson-Boltzmann predictions is observed when the potential is sufficiently negative to make the surface degenerate. An important point here is that all the experimental points lie, if anything, *below* the Poisson-Boltzmann curves. The presence of surface-states would cause departures in the opposite direction. A second feature of the data is that at sufficiently negative bias (~ -0.4 volt with respect to the flat-band potential) the capacitance becomes almost completely independent of the bulk electron density and appears to approach a limiting value of the order of 6 μf per cm^2 at the most negative potentials studied.

A large number of factors neglected in the derivation of (10) can operate to make the measured capacitance lower than the Poisson-Boltzmann values. For one thing, the capacitances of the Helmholtz and Gouy layers are in series with the space-charge capacitance, and it may be incorrect to use the value of 30 μf per cm^2 for C_H in obtaining the space-charge capacitance from the measured capacitance. This value was taken from work with the mercury electrode. Since the highest capacitance measured was about 6 μf per cm^2, it seems unlikely that this is the cause of the discrepancy. In this connection it should be noted that, if the surfaces were rough, any error in using too large a value for C_H would be partly compensated.

The most important effect causing the capacitance to fall below the Poisson-Boltzmann prediction is the onset of degeneracy at the surface. When the bands bend down sufficiently to make the Fermi level lie within the conduction band, the effective density-of-states treatment breaks down. Since only one electron can be put into each level and the number of levels at each energy is finite, the conduction-band electrons at the surface must have energies lying well into the conduction band. This counteracts the tendency of the electric field to pull electrons towards the surface, and makes the electron density at the surface less than that predicted from the classical treatment. In consequence, the capacitance is also less than classical.

* Degeneracy starts to be important when the Fermi level lies above the conduction band.

The effects of surface degeneracy may be treated quantitatively by the use of Fermi-Dirac statistics in conjunction with an effective mass approximation.[19] The electron density n at any point in the space-charge layer is then given by

$$n = \frac{2}{\sqrt{\pi}} N_c \left(\frac{m^{(N)}}{m}\right)^{\frac{3}{2}} \int_0^\infty \frac{\epsilon^{\frac{1}{2}}\, d\epsilon}{1 + e^{\epsilon+\varphi-y}}, \tag{15}$$

where N_c is the effective density-of-states [see (7)], $m^{(N)}$ is the density-of states mass and φ is the Fermi energy measured from the conduction band in the bulk of the crystal. ϵ and φ are in units of kT. When this expression is inserted in the Poisson equation, one may approximate the resulting integrals (within a few per cent or less if $Y - \varphi > 3$) and obtain the capacitance in the form.

$$C_{sc} = \left\{\frac{5\kappa\epsilon_0 N_c q^2}{3\pi^{\frac{1}{2}} kT}\left[\frac{m^{(N)}}{m}\right]^{\frac{3}{2}}\right\}^{\frac{1}{2}} (Y - \varphi)^{\frac{1}{4}}. \tag{16}$$

Using Hutson's[12] values for $m^{(N)}$ and κ, this becomes, at room temperature,

$$C_{sc} = 6.1(Y - \varphi)^{\frac{1}{4}} \qquad \mu\text{f per cm}^2. \tag{17}$$

The capacitance in this degenerate region depends upon bias only as the one-fourth power and depends almost negligibly upon the donor density (through φ). As the curves in Fig. 8 marked "Poisson-Fermi" show, the qualitative features of (17) are reasonably well in accord with the experimental result. This is particularly striking in view of the fact that no additional adjustable parameters are used in drawing the curves.

It is tempting to try to rationalize the quantitative discrepancy between the experimental data and (17). It could be due to using too high a value for C_H. Alternatively, the use of Hutson's value of 0.5 for the density-of-states mass, $m^{(N)}$, may not be valid, since we are concerned here with the density-of-states at the *surface* and Hutson's measurement was for the bulk. If anything the density of states near the surface should be less than that in the bulk. If the density-of-states mass near the surface were half the bulk value, the agreement between theory and experiment would be truly quantitative.

A more exact treatment than the Poisson-Fermi equation, presumably treating the electrons as quantum particles and including discreteness-of-charge effects, is probably required before any sound conclusions can be drawn from the discrepancy.

4.3 *Results with Boron and Hydrogen Donors*

The results presented above are typical of the majority of crystals that have been studied. Since they agree quite well with the simple

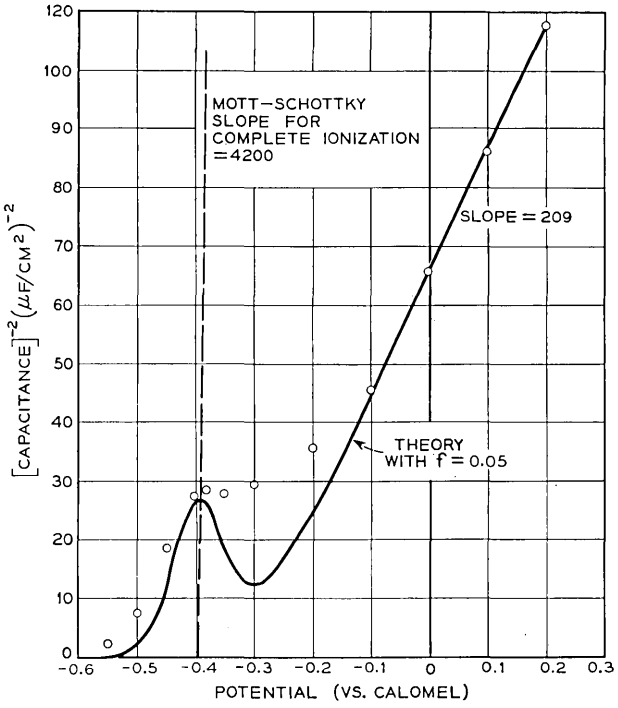


Fig. 9 — Mott-Schottky plot for a crystal containing boron, a low-lying donor; the dotted line shows the slope expected at this electron density if all the donors were ionized, solid line is the complete theory [equation (10)] if the donors are 5 per cent ionized.

theory outlined above, we are reasonably confident that they represent the "normal" behavior for crystals with a uniformly distributed immobile donor lying reasonably close to the conduction band. A number of "anomalous" types of behavior have also been observed. As we shall see, all but one of these anomalous results are understandable, in at least a qualitative fashion, as arising from donor inhomogeneity, donor mobility and/or weak donor ionization. They are presented here in the interest of completeness and also to illustrate the use of surface capacitance measurements in studying the bulk properties of semiconductors.

Fig. 9 shows the behavior typical of one lot of crystals into which boron was inadvertently introduced.* Here we plot $1/C^2$ versus bias as in Fig. 3. For positive potentials (versus calomel) the data were nicely linear as prediced by (12). However, the slope was only about one-twentieth of the value expected if all the donors were ionized, when the

* It was subsequently identified by spectral analysis.

conductivity and Hall effect results, indicating an electron density of about 3.9×10^{15}, would predict a Mott-Schottky slope of 4200 (μf per cm^2)$^{-2}$ volt^{-1}. This is shown by the dotted line in the figure. But the observed slope is only 209. Conductivity and capacitance measurements made after successive etching in H_3PO_4 indicated that the crystal was essentially uniform in composition (within 20 per cent or so) and ruled out the possibility of an excessive donor concentration near the surface. The simplest explanation of the much reduced Mott-Schottky slope in Fig. 9 is that the donors in the bulk of the crystal are only about one-twentieth ionized.

The dependence of capacitance on bias in the region of negative potentials is complex for this crystal. It confirms in at least a qualitative way the assumption of incomplete ionization. With increasingly negative bias, the theoretical capacitance computed from (10) goes through a maximum and then a minimum (located near the flat-band potential) before finally rising exponentially as predicted by (13).

The solid curve in Fig. 9 is computed for $f = 0.05$, this value being obtained from the slope at strongly positive potentials. As can be seen, the experimental data show behavior that is qualitatively similar to theory. The value of the capacitance at the minimum and also the potential at which it occurs are in good accord with the theory. The only discrepancy is that the capacitance maximum (the minimum in the curve as plotted) is very much shallower than predicted by (10). The reason for this is not presently understood. It might conceivably arise from the presence of two kinds of donors. However, curve fitting for such a situation, with the consequent introduction of two more parameters, does not seem waranted.

If one accepts the value of 0.05 for f, (6) allows one to compute the energy of the boron level in zinc oxide. Assuming a donor degeneracy of 2 and Hutson's value of 0.5 m for $m^{(N)}$, one obtains a value of 0.30 ev for E_D. This agrees quite well with a preliminary measurement by Zetterstrom[20] of the temperature dependence of Hall effect and conductivity which gave a value of 0.32 ev for E_D.

Hydrogen-doped crystals showed complicated frequency effects and long-term transients.

V. SUMMARY

i. The differential capacitance of single-crystal zinc oxide electrodes has been studies as a function of bias, frequency, surface treatment and crystal impurity content in the range from 10^{15} to 10^{19} ions per cc.

ii. Under nondegenerate surface conditions the results on "as-grown" and indium-doped crystals are quantitatively in accord with the capacitance calculated from the Poisson-Boltzmann equation, surface states playing no role whatsoever. Boron-doped crystals can also be understood in these terms if one assumes boron to be a low-lying donor.

iii. The variation of the flat-band potential with surface treatment indicates that a variable surface dipole is present. A strongly anomalous dependence of flat-band potential on donor density is observed at densities around 2×10^{18} donors per cc. It is speculated that this may be a direct measure of an anomalous variation in Fermi energy.

iv. Under conditions where the surface is degenerate, the results are in almost quantitative agreement with those predicted by the Poisson-Fermi equation.

VI. ACKNOWLEDGMENTS

I am indebted to J. J. Lander and A. R. Hutson for many stimulating discussions of this work. I am also indebted to R. T. Lynch, who grew most of the crystals used. I am particularly appreciative of the efforts of R. B. Zetterstrom, who made all of the Hall-effect measurements.

REFERENCES

1. Brattain, W. H. and Garrett, C. G. B., B.S.T.J., **34,** 1955, p. 129.
2. Dewald, J. F., in Hannay, B. N., ed., *Semiconductors,* Reinhold Publishing Corp., New York, 1959, p. 727.
3. Green, M., in Bockris, J. O., ed., *Modern Aspects of Electrochemistry, No. 2,* Academic Press, New York, 1959.
4. Bohnenkamp, K. and Engell, H. J., Z. Elektrochem., **61,** 1958, p. 1184.
5. Parsons, R., in Bockris, J. O., ed., *Modern Aspects of Electrochemistry, No. 1,* Academic Press, New York, 1954.
6. Bardeen, J., Phys. Rev., **71,** 1947, p. 717.
7. Law, J. T., in Hannay, N. B., ed., *Semiconductors,* Reinhold Publishing Corp., New York, 1959, p. 676.
8. Schottky, W., Z. Physik, **113,** 1939, p. 367; **118,** 1942, p. 539.
9. Mott, N. F., Proc. Royal Soc., **A171,** 1939, p. 27.
10. Garrett, C. G. B. and Brattain, W. H., Phys. Rev., **99,** 1955, p. 376.
11. Macdonald, J. R., J. Chem. Phys., **29,** 1958, p. 1346.
12. Hutson, A. R., Phys. Rev., **108,** 1957, p. 222.
13. Scharowsky, E., Z. Physik, **135,** 1953, p. 318.
14. Thomas, D. G. and Lander, J. J., J. Chem. Phys., **25,** 1956, p. 1136.
15. Thomas, D. G., J. Phys. Chem. Solids, **9,** 1958, p. 31.
16. Dewald, J. F., Abst. New York Meeting, Amer. Electrochem. Soc., 1958.
17. Hutson, A. R., private communication.
18. Grahame, D. C., J. Am. Chem. Soc., **68,** 1946, p. 301.
19. Hannay, N. B., *Semiconductors,* Reinhold Publishing Corp., New York, 1959, p. 13.
20. Zetterstrom, R. B., private communication.

The Square of a Tree

By IAN C. ROSS and FRANK HARARY*

(Manuscript received June 4, 1959)

The adjacency matrix of a graph of n points is the square matrix of order n, in which the i,j element is one if and only if the ith point and the jth point are adjacent, or $i = j$; and is zero otherwise. Let A be the adjacency matrix of graph G considered as a boolean matrix so that $1 + 1 = 1$. Then G^2, the square of G, is the graph whose adjacency matrix is A^2. We obtain a necessary and sufficient condition for a graph to be the square of a tree by providing an algorithm for determining a tree that is the square root of any graph known to be the square of some tree. This algorithm cannot be carried through when a graph is not the square of a tree. It is shown that, if a graph is the square of a tree, then it has a unique tree square root. The method utilizes a previous result for determining all the cliques in a given graph, where a clique is a maximal complete subgraph. This result was obtained while attempting the more general problem of characterizing boolean matrices having a square root, or, in general, an nth root.

I. INTRODUCTION

The correspondence between graphs, matrices and relations is well known and presents an interesting field of investigation. With any graph G there is associated a symmetric square matrix of 0's and 1's, called its "adjacency matrix". Forming the boolean square of this matrix, we may call the corresponding graph "the square of G". It is an open problem (communicated to us by N. J. Fine) to characterize those graphs that have at least one square root graph, and in general those graphs that have an nth root for any positive integer n. This paper presents a partial solution to the general problem by characterizing those graphs with a tree for a square root. It is also shown that, if the graphs obtained on squaring two trees are isomorphic, then the trees themselves are isomorphic.

In the course of the development of this paper, an algorithm is given

* Princeton University and Institute for Advanced Study (presently at University of Michigan); work supported by a grant from the Office of Naval Research to the Princeton University Logistics Project.

for constructing a tree from a given graph known (or found by the characterization) to be the square of a tree. This is based on a previous method for determining all the cliques in any graph.[1]

A graph G consists of a finite collection of *points* together with certain *lines* joining pairs of distinct points. Two points of G are *adjacent* if they are joined by a line. A *complete graph* is one in which every two distinct points are adjacent. A *clique* is a maximal complete subgraph of G containing at least three points. A point of G is *cliqual* if it is in at least one clique; it is *unicliqual* if it is in exactly one clique and *multicliqual* if it is in more than one clique. Two points are *cocliqual* if there is a clique containing both of them. The *neighborhood* of a point b of G consists of b together with all points adjacent to b. A point is *neighborly* if it is cocliqual with each point in its neighborhood. Thus, every neighborly point is cliqual.

A *path* is a collection of lines of the form b_1b_2 , b_2b_3 , $\cdots$, $b_{t-1}b_t$, where these t points are distinct. A *cycle* consists of a path together with the line b_tb_1 joining its end points. A graph is *connected* if there is a path between any two points. A *tree* is a connected graph with no cycles. An *endpoint* of a graph is one which is incident to exactly one line. It is well known[2] that every tree containing more than one point has an endpoint. A *next-point* is one which is adjacent to an endpoint. An *articulation point* or *cut point* of a connected graph is one whose removal results in a disconnected graph. A *block* is a connected graph with no articulation points.

Let the points of G be b_1 , $\cdots$, b_p . The *adjacency matrix* of G is the matrix $A = (a_{ij})$, where $a_{ij} = 1$ if points b_i and b_j are adjacent and $a_{ij} = 0$ otherwise, except that we take (arbitrarily) the diagonal of A to contain only 1's. By the *boolean product* of two matrices of 0's and 1's is meant their ordinary product with the stipulation that $1 + 1 = 1$. The *square of* G, written G^2, is that graph whose adjacency matrix is A^2 (all matrix products will be regarded as boolean).

II. LEMMAS

It is convenient to derive the following sequence of lemmas. In this section, we have as the hypothesis that G is a given graph known to be the square of some tree T. The tree T itself is not given; only the graph G is available and it is assumed that G has at least three points.

Lemma 1: G is complete if and only if T has exactly one next-point.

Proof: If T has exactly one next-point b, then the neighborhood of b contains all the points of T. Therefore G has exactly one clique and is complete. If G is complete, then T must have a unique next-point. For

if T has two distinct next-points, then there are two distinct points of G that are endpoints of T and are not adjacent in G.

Lemma 2: Every point of G is neighborly and G is connected.

Proof: Since $G = T^2$, every point of G lies on a triangle and so is cliqual. Further, every point is neighborly since every line of G is contained in a triangle. Obviously, the square of a connected graph contains that graph and is defined on the same set of points; G is therefore connected.

Lemma 3: Every clique of G is the neighborhood of a nonendpoint of T, and conversely.

Proof: We first remark that the neighborhood of an endpoint can be a clique only in a graph of two points that we exclude. Obviously, the neighborhood of any nonendpoint b of T generates a complete subgraph of G. It remains to show that this subgraph is maximal. This is immediate, since any point not in the neighborhood of b cannot be adjacent in G to all points of the neighborhood.

To prove the converse, let C be a clique of G. If an endpoint b of T is in C, then the clique C must consist of the neighborhood of the next-point of b. For any point of G not in this neighborhood cannot be adjacent to b in G. If C does not contain an endpoint of T, let c be any point of C. If C is the neighborhood of c, there is nothing to prove. Otherwise, there is a point d of C not in the neighborhood of c. In this case, both points c and d must have some common adjacent point e. Thus C is the neighborhood of e.

Lemma 4: If G has more than one clique, then b is a unicliqual point of G if and only if b is an endpoint of T. Alternatively, we may say that if G is not complete then b is a multicliqual point of G if and only if it is a nonendpoint of T.

Proof: If b is an endpoint of T, then the only clique containing b is given by the neighborhood of the next-point of b. If b is unicliqual in G, then it cannot be adjacent to two distinct points of T. For b would then belong to more than one clique, since by hypothesis G contains more than one clique.

Lemma 5: Two nonendpoints of T are adjacent in T if and only if their neighborhoods are cliques of G that intersect in exactly two points.

Proof: If two nonendpoints of T are adjacent in T, then they are both contained in the cliques in G given by their neighborhoods in T in accordance with Lemma 3. These two nonendpoints cannot both be adjacent in T to a common third point since T has no cycles. Hence the intersection of these two neighborhoods contains exactly these two points.

Conversely, if two distinct cliques of G intersect in exactly two points, then the corresponding nonendpoints in T (in accordance with Lemma 3) are adjacent in T. For if these two points are not adjacent, their neighborhoods in T can intersect in at most one point since T is a tree and has no cycles.

Lemma 6: G is a block.

Proof: Lemma 2 shows that G is connected. Points that are adjacent in T are also adjacent in G. Two points both adjacent to any given point in T are adjacent to each other in G. Therefore the removal of any point of G cannot disconnect it.

III. CHARACTERIZATION

We now state conditions that we shall see characterize every graph that is the square of a tree and demonstrate that, if $G = T^2$, then G satisfies these conditions. In the next section we shall present an algorithm for finding a tree root of such a graph, and shall show that the algorithm can be applied to graphs that meet the conditions of the present section. Therefore, a graph will be shown to meet these conditions if and only if it has a square root that is a tree.

The characterization has two cases. In Case 1, $G = K_p$, the complete graph of p points; in case 2, $G \neq K_p$.

Case 1

Consider the tree consisting of one point joined with all the others. When this tree is squared, the result is the complete graph. We illustrate with Fig. 1, in which $T^2 = K_5$.

Case 2

Consider the following five conditions for a graph $G \neq K_p$:

i. Every point of G is neighborly and G is connected.

ii. If two cliques meet at only one point b, then there is a third clique with which they share b and exactly one other point.

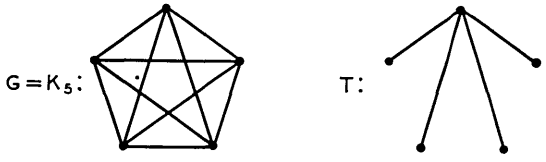


Fig. 1 — Graphs for Case 1.

iii. There is a one-to-one correspondence between the cliques and the multicliqual points b of G such that the clique $C(b)$ corresponding to b contains exactly as many multicliqual points as the number of cliques which include b.

iv. No two cliques intersect in more than two points.

v. The number of pairs of cliques that meet in two points is one less than the number of cliques.

We now prove that if there is a tree such that $T^2 = G$, then G will either be complete or will satisfy the above five conditions. We have already settled Case 1 in Lemma 1 and Fig. 1. We now turn to Case 2, where $G \neq K_p$.

The necessity of the first four of these conditions for a graph $G \neq K_p$ that is the square of a tree follows readily from the lemmas. Condition i is a restatement of Lemma 2. Condition ii follows from Lemmas 3 and 5, for, if $C_1 \cap C_2 = \{b\}$, then neither C_1 nor C_2 can be the clique associated with point b in accordance with Lemma 3. Hence, there is a third clique C_3 corresponding to b, and C_3 has the relationship stated in ii with C_1 and C_2 by Lemma 5. Condition iii is a combination of Lemmas 3 and 4, while iv is implied by Lemma 5. Condition v follows immediately from the well-known theorem that for trees the number of points exceeds the number of lines by one (see p. 51, satz 9 of Ref. 2).

IV. ALGORITHM AND THEOREM

If the graph G is complete, we have already seen that there is a unique tree T such that $G = T^2$, and we have illustrated this with Fig. 1.

Now, let G be a graph that is not complete and satisfies the five conditions of Section III. We now show how to construct a unique tree T (up to isomorphism) that is a square root of G.

Step 1

Find all the cliques of G in accordance with the method of Ref. 1.

Step 2

Let the cliques of G be $C_1, \cdots, C_n$. Then $n > 1$, since G is not complete and condition i holds. Consider a collection of multicliqual points $b_1, \cdots, b_n$ corresponding to these cliques in accordance with condition iii. These are to be the nonendpoints of the tree T being constructed. Find all the pairwise intersections of the n cliques. By condition iv, no cliques meet in more than two points. We now form a graph S by joining the points b_i and b_j by a line if and only if the corresponding cliques C_i and C_j intersect in two points. By condition v, S is a tree.

Step 3

For each clique C_i of G, let n_i be the number of unicliqual points. Thus n_i is a nonnegative integer. To the tree S obtained in Step 2 attach n_i endpoints to b_i, obtaining a tree T. Since the points of G are either multicliqual or unicliqual and the points of T are either nonendpoints or endpoints, the number of points of G and T is equal.

It is clear that the tree T constructed by this algorithm is a tree square root of G. It also follows that the five conditions of the preceding section are sufficient for an incomplete graph to be the square of a tree. Further, the algorithm results in a unique tree T, up to isomorphism.

We have just proved the following theorem about the relationship between trees and their squares:

Theorem: Let T_1 and T_2 be trees. If T_1^2 and T_2^2 are isomorphic, then T_1 and T_2 are isomorphic.

V. EXAMPLE

We illustrate the algorithm with the graph G of Fig. 2, which meets the conditions for being the square of a tree. In this graph we indicate the points for convenience by the numerals 1, . . . , 8.

By the method of Ref. 1 we find all the cliques of G. There are four cliques whose composition is as follows:

C_1: 1256,
C_2: 1234,
C_3: 237,
C_4: 248.

In accordance with Step 2 of the algorithm, we introduce the points b_1, b_2, b_3 and b_4. Among the four cliques there are pairwise intersections between C_2 and each of the other cliques, but the other cliques have only single point intersections with each other. We therefore obtain the graph S shown in Fig. 3, which is a subtree of the tree T under construction.

We now proceed in accordance with Step 3. Let U_i be the set of unicliqual points of C_i; then we have:

U_1: 56,
U_2: —,
U_3: 7,
U_4: 8.

On joining each point b_i to the number of points in the set U_i we obtain the tree T shown in Fig. 4.

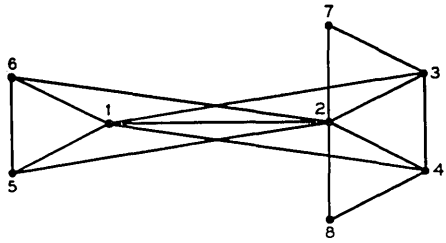


Fig. 2 — Graph G for the example.

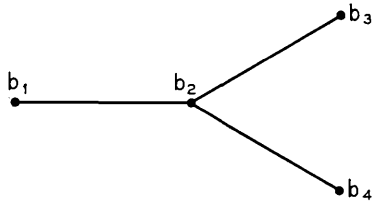


Fig. 3 — Graph S for the example.

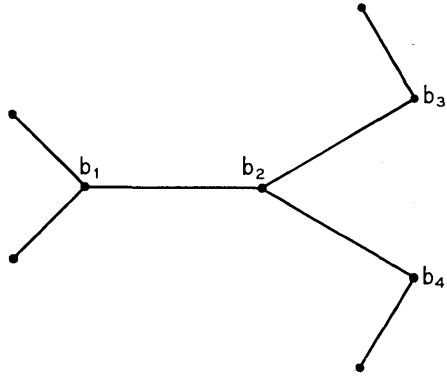


Fig. 4 — Graph T for the example.

REFERENCES

1. Harary, F. and Ross, I. C., A Procedure for Clique Detection Using the Group Matrix, Sociometry, **20**, 1957, p. 205.
2. König, D., *Theorie der endlichen und unendlichen Graphen*, Academische Verlagsgesellschaft, Leipzig, 1936.

Theory of a Frequency-Synthesizing Network

By B. M. WOJCIECHOWSKI

(Manuscript received December 28, 1959)

The theoretical basis for designing frequency-combining and selecting circuits is developed. By the introduction of "sideband algebra" and of a frequency symbolic network, the new method offers formal design procedures in place of intuitive ones. This leads directly to finding optimal solutions for frequency-adding or frequency-subtracting problems without limitations as to the relative frequency ratios. The derivation of typical frequency-synthesizing circuits, such as "slave" oscillators and digital frequency selection systems, is discussed, and examples of practical solutions are given.

I. THE SYMBOLIC FREQUENCY-COMBINING NETWORK

1.1 *Introduction*

The frequency accuracy of ac signal sources plays an increasingly important role in various technical and physical fields of precision measurements.[1] A primary reason for this perhaps is a numerical property associated with signal frequencies that is rather unique among quantitative physical phenomena: with the use of relatively simple nonlinear elements[2] and electric wave filters,[3] signal frequencies, like numerical quantities, can be both added to or subtracted from each other and multiplied or divided by an integer.[4,5] The digital character of these operations permits the devising of systems that provide extremely high frequency accuracies and almost unlimited resolution capabilities.[6,7]

In many of these techniques related to, and based on, signal-frequency generation, adjustability, readability or interpolation, the combining of two signal frequencies to obtain a single frequency signal in the form of their sum or difference is one of the most basic operations. When the ratio between the two frequencies to be thus combined is low, the method is quite straightforward, since the desired result can be arrived at by using a nonlinear element (modulator) and an appropriate electrical

wave filter. However, when the ratio between the two frequencies increases, the problem of separating the desired sideband from the undesired products becomes progressively difficult and, eventually, impractical.

Another class of frequency-combining problems where a satisfactory solution is particularly difficult to find consists of cases when two signal frequency sources must track each other by a constant, relatively small frequency interval. A classic solution to this problem uses a "slave" oscillator system that employs both mechanical and electrical servomechanism control elements.[8] The servomechanisms perform satisfactorily when the "master" frequency is not subject to rapid frequency variations and random drifts, or when tracking accuracy requirements are not critical. With a high rate of these variations, however, the inherent mechanical and electrical inertia of these systems produces unacceptable tracking errors. Thus, a solution based on a noninertial system is necessary when high-accuracy requirements must be met.

The theory presented here leads to a general method of designing frequency-combining systems having no limitations as to the relative frequency ratios and employing only standard circuit elements such as modulators, filters and oscillators. This method can also be extended to variable frequency sources and to the design of "slave" oscillators and signal frequency systems covering extensive ranges that may be adjustable by decade or other digital steps.

1.2 *Sideband Algebra of Frequency-Combining Circuits*

Operations with numbers representing or standing for signal frequencies are restricted in certain ways by the very physical nature of the frequency-combining means and methods. In order to facilitate the process of logical deductions leading to the solutions of frequency-combining problems, in the following discussion these restrictions are codified and then imposed upon the algebraic operations that are directly related to the combining of signal frequencies. The rules thus evolved are called *sideband algebra.*

1.2.1 *An Elemental Frequency-Combining Scheme*

As is generally known, two signal frequency sources, f_1 and f_2, applied to a nonlinear element (modulator) produce a spectrum of modulation products of the general form $pf_2 \pm qf_1$, where p and q are integers and "+" or "−" indicates the upper or the lower sideband spectra, respectively.[9] For the present considerations, the most important among these

products are the upper ($f_2 + f_1$) and the lower ($f_2 - f_1$) single-frequency sideband products and the carrier frequency (the higher of the two primary frequencies). In order to obtain the desired combination of the two frequencies, all the other modulation products need to be suppressed to the desired degree. In the circuits under consideration, this suppression is accomplished by an electrical wave filter.

An elemental frequency-combining scheme is shown in Fig. 1(a). Two signal-frequency sources, f_1 and f_2, are connected to a nonlinear element, or modulator, M, while the electrical wave filter, F, passes the desired sum or difference frequency, f_c, and suppresses to a desirable degree the other modulation products, in particular the other sideband product and the carrier frequency f_2. This operation may be expressed by the formula

$$f_2 +/- f_1 = f_c \tag{1}$$

where the symbol $+/-$ stands for "either the upper or the lower sideband product".

It can be noted that the frequency differences between a desired modulation product, f_c, and the nearest unwanted modulation products are equal to f_1, or the lower of the two frequencies to be combined. Therefore, the lower the relative value of the f_1 frequency (or the higher the ratio between f_2 and f_1 frequencies), the more difficult are the frequency-discrimination requirements that must be met by the filter design. Thus, if n denotes the maximum frequency ratio that can be accommodated efficiently within a practical filter design under consideration, then two numerical quantities representing frequencies f_1 and f_2 can be directly added to or subtracted from each other only if

$$\frac{f_2}{f_1} \leqq n, \tag{2}$$

where f_1 is the lower of the two frequencies.

In Fig. 1(a) all three frequencies involved in the elemental frequency-

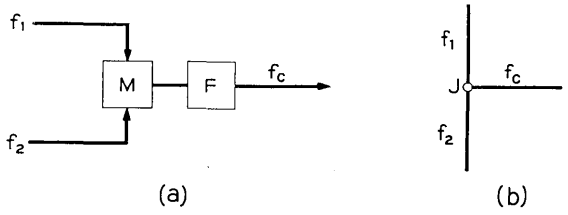


Fig. 1 — Elemental frequency-combining scheme.

combining scheme are known, and also it is predetermined which of these three frequencies, available at the filter output, is the result of combining the other two. It should be noted, however, that, with the same three frequencies involved, two other modulator-filter arrangements are possible, where either of the two frequencies shown in Fig. 1(a) as the "input" frequencies can be obtained as a combination of the two other frequencies. To contain all of these possibilities, a symbolic frequency-combining diagram, as shown in Fig. 1(b), is introduced. Here the sequential modulator-filter arrangement of Fig. 1(a) is replaced by an asequential arrangement of a "junction point," J, at which the three radial lines, each representing signal frequency, converge. In order that the symbolic frequency-combining diagram have a physical basis and be translatable, when needed, into a practical arrangement as shown on Fig. 1(a), the following self-explanatory rules will apply:

i. The highest of the three frequencies converging at the junction point is equal to the sum of the two other frequencies.

ii. If the highest of the three frequencies is the derived output frequency, then the given junction point performs summation (selection of the upper sideband frequency combination); if the highest frequency is one of the two original input frequencies, then the given junction point performs subtraction (selection of the lower sideband frequency combination).

1.2.2 *The Frequency-Ratio Index, k*

Let us assume that the ratio between two frequencies to be combined, f_k and f_0, is considerably larger than an acceptable n ratio, as defined in (2). In this case, the ratio f_k/f_0 can be presented as equal to, or smaller than, the product of a minimum number, k, of individual n ratios, so that:

$$\frac{f_k}{f_0} \leqq n_1 n_2 \cdots n_k ,$$

where n_1, n_2, $\cdots$, n_k represent individual frequency ratios, as defined by (2) for a sequence of frequency levels between the f_0 and f_k frequencies. The integer k will thus represent the lowest number of realizable modulation levels that must intervene between the f_0 and f_k frequencies in order to keep the frequency-combining system within the requirements of a practical filter design. If we assume that n' represents the geometric mean of the n_1, n_2, $\cdots$, n_k ratios, or

$$n' = \sqrt[k]{n_1 n_2 \cdots n_k},$$

then

$$\frac{f_k}{f_0} \leqq (n')^k. \tag{3}$$

The k integer, called the *frequency ratio index*, is a basic design parameter of the frequency-combining systems described here.

Thus, to make possible a transition from the low frequency, f_0, to the high frequency, f_k, a number of intermediate frequency sources, f_a, must be introduced. In an arrangement such as shown in Fig. 2, let f_0 and f_k represent two high-ratio frequencies to be combined. Similarly, $J_1, J_2, \cdots, J_k$ junction points represent modulator-filter stages, while $f_{a1}, f_{a2}, \cdots, f_{a(k-1)}$ represent intermediate frequencies. Inspection of the arrangement shown in Fig. 2 indicates that the number, m_a, of these intermediate frequency sources is:

$$m_a = k - 1. \tag{4}$$

It should be noted that the straightforward modulation scheme, as shown in Fig. 2, provides a single-sideband combination, not only of the desired f_0 and f_k frequencies, but also of all the intermediate frequencies. This scheme, therefore, while illustrating a physically realizable transi-

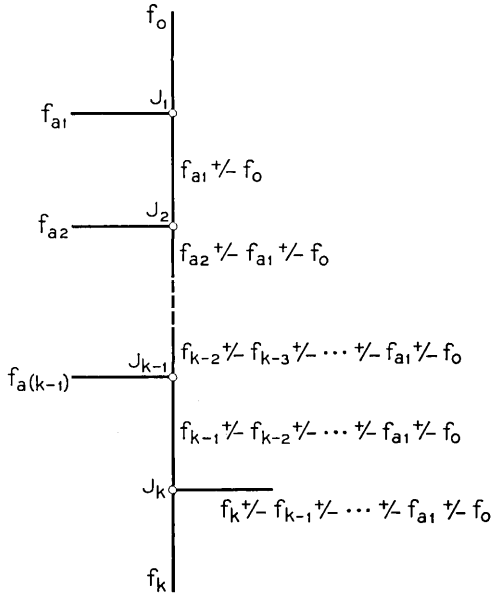


Fig. 2 — Straightforward modulation scheme.

tion between two high-ratio frequency levels, does not provide a solution of the problem of obtaining the sum or difference of the two original high-ratio frequencies alone.

1.2.3 *Algebraic Signs of Frequency Combining*

Denotations of frequency, expressing, in general, a number of events per unit of time, usually have no algebraic "positive" or "negative" signs ascribed. In the present considerations, however, we are not only concerned with frequencies as nominal "numbers of events," but also with their inherent variations with time, generally known as *drifts*. The individual and nonreproducible character of these variations, as well as the necessity of eliminating some of them along with the associated frequency components, leads eventually to certain algebraic methods of manipulating signal frequencies. These manipulations may be facilitated if, within the meanings and conditions described below, algebraic signs are assigned to the symbols representing frequencies within the frame of the sideband algebra.

It can be observed that the instantaneous value of the combined drift of a frequency that is derived as an upper sideband combination of two frequencies is equal to the sum of the instantaneous values of the individual drifts of the component frequencies. On the other hand, the instantaneous value of the drift of a lower sideband frequency combination is equal to the difference of the individual drifts. Thus, the physical meaning of selecting the upper or the lower sideband combinations can be translated into algebraic concepts associated with positive or negative signs, respectively. In relation to these operations, the following rules concerning the algebraic frequency signs are proposed:

i. It is assumed that the frequency of any primary signal source (such as an oscillator, for instance), before being combined within the system, has a positive sign.

ii. When combining two single frequencies (each of which may be the result of some antecedent combining operations), the higher of the two frequencies (with its original component frequencies, if any) does not change its sign; the lower of the two frequencies (with its original component frequencies, if any) retains its sign in the upper sideband combination, but reverses it in the lower sideband combination.

This completes a set of rules instrumental for the derivation of the frequency-combining diagram as shown below. By application of the above rules, it will be proved also that this diagram provides a general solution for the problem of adding or subtracting two frequencies independently of their ratio.

1.3 *The General Frequency-Combining Network*

Let us consider a symbolic network such as the one shown in Fig. 3. This network consists of a number of junction points and branches that link the adjacent points. As explained previously, each of the junction points represents symbolically a realizable modulator-filter ensemble, while each of the branches represents a single signal frequency. Besides the frequencies represented by bilateral network branches (such as f_1, f_2, f_{01}, etc.), there are three more frequencies (f_0, f_k and f_{0k}), each represented by a line unilaterally connected to one of the three corners of the network. The first two of these frequencies are the originally-to-be-combined frequencies, f_k and f_0, respectively (marked by incoming arrows), while the third is the output frequency from the modulation network (marked by an outgoing arrow). Each of the three frequencies that converge at any junction-point must meet the requirements of (1) and, in addition, combinations of two of these frequencies must meet the requirements of (2).

Each of the bilateral network branches may represent either an "auxiliary" frequency applied simultaneously to two modulator-filter

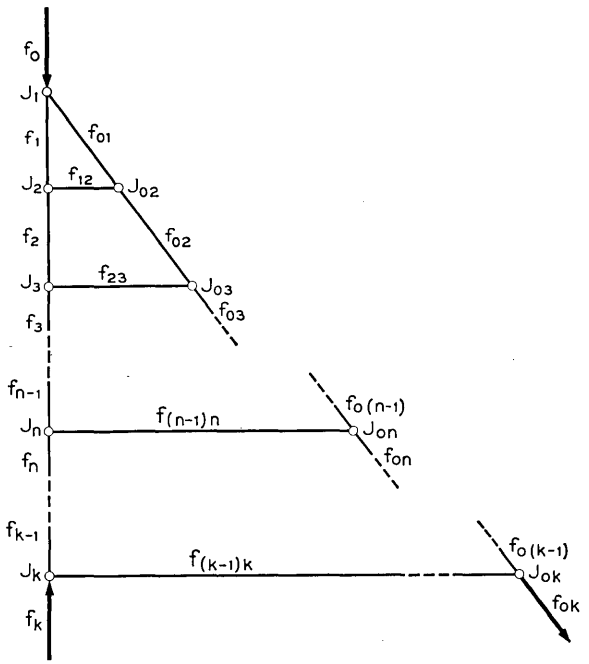


Fig. 3 — Symbolic frequency-combining network.

ensembles from an extraneous signal source, or it may represent a derived frequency that is a single sideband combination of one or both of the original frequencies and of other auxiliary frequencies. This method of connecting bilaterally all auxiliary frequencies and their combinations between a pair of adjacent junction points is significant, since it permits eventual cancellation of all auxiliary component frequencies and their drifts from the final network output. Conversely, the two original frequencies, f_0 and f_k, each connected only to a single junction point (J_1 and J_k, respectively) and thus represented by unilateral lines, are the sole "surviving" components whose single sideband combination is present at the eventual output from the frequency-combining system under discussion.

1.3.1 *Modulation Frequency Levels and Frequency-Network Meshes*

It has been previously established that the number of modulation levels intervening between the low (f_0) and the high (f_k) ratio frequencies to be combined is equal to k [see (3)]. As shown in Fig. 3, at the lowest ("first") frequency level there is only one modulator-filter ensemble, which is shown in the symbolic diagram as junction point J_1, to which point the lower frequency to be combined, f_0, is applied. At each of the following modulation levels, however, there are two junction points (J_2–J_{02}, J_3–J_{03}, etc.). Eventually, to one of the two junction points at the highest frequency level, the higher of the two frequencies to be combined, f_k, is applied, while the above-mentioned output frequency f_{0k} is taken from the other junction point, J_{0k}.

The total number of junction points can be found from Fig. 3 as:

$$m_j = 2k - 1. \tag{5}$$

The described configuration of junction points and interconnecting links forms a certain number of meshes within the frequency-combining network. The first of these meshes contains but three branches, while each of the succeeding meshes has four branches. Any two adjacent meshes are coupled with each other by two common junction points and by the common branch linking these points.

On the basis of the foregoing considerations, it can be found that the conditions sufficient for determining all frequencies within any mesh, are as follows:

i. The out-of-mesh frequencies, applied unilaterally to all mesh corners except one, must represent input frequencies of known values.

ii. Within any given mesh one branch must represent a known

extraneously introduced "auxiliary" frequency that is simultaneously (i.e. bilaterally) applied to two adjacent junction points of the mesh.

iii. A branch that is common to two meshes may represent an auxiliary frequency for either one of two meshes, but not for both of them.

Application of these rules will be illustrated in the following discussion.

1.3.2 *Frequency Relationship Within the Symbolic Network*

Following the above rules, we stipulate that, within each mesh of the network shown in Fig. 3, one branch represents an auxiliary frequency assigned to this mesh. The number of the auxiliary frequencies within the network will be thus equal to the number of meshes, or:

$$m_m = k - 1. \tag{4'}$$

It should be noted that this is the same number of auxiliary frequencies as is found in (4) in relation to Fig. 2.

For the sake of simplicity, let us assume that all auxiliary frequencies, f_a, are located in the series-connected network branches marked in Fig. 3 as $f_1, f_2, \cdots, f_{k-1}$. If $n_1, n_2, \cdots, n_k$ denote permissible frequency ratios as defined in (2) for $J_1, J_2, \cdots, J_k$ junction points, respectively, then

$$\frac{f_1}{f_0} = n_1; \quad \frac{f_2}{f_1} = n_2; \quad \cdots; \quad \frac{f_k}{f_{k-1}} = n_k. \tag{2.1}$$

Each of the third frequencies $f_{12}, f_{23}, \cdots, f_{(k-1)k}$ converging at these junction points will have values complementary to two other frequencies. Thus,

$$\text{for } J_2: \quad f_{12} = f_2 +/- f_1; \tag{1.1}$$

$$\text{for } J_3: \quad f_{23} = f_3 +/- f_2; \tag{1.2}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\text{for } J_k: \quad f_{(k-1)k} = f_k +/- f_{k-1}. \tag{1.k}$$

Similarly, the third frequency converging at the J_1 junction point can be found as

$$f_{01} = f_1 +/- f_0. \tag{1.1.1}$$

Examination of (1.1) and (1.1.1) shows that both frequencies f_{12} and f_{01} contain the common frequency component, f_1. From (1.1), (2.1) and

(1.1.1) it can be found that

$$\frac{f_{12}}{f_{01}} = n_1\left(\frac{n_2 \pm 1}{n_1 \pm 1}\right).$$

In cases of practical filter design, values of n_1, n_2, $\cdots$, n_k can be assumed to be considerably larger than one. Thus,

$$\frac{f_{12}}{f_{01}} \cong n_2. \tag{6}$$

From (6) it can be concluded that frequencies f_{12} and f_{01} can be combined at the junction point J_{02} by using a filter of a design similar to the filter at J_2. By doing so, we obtain:

$$f_{02} = f_{12} +/- f_{01} = (f_2 +/- f_1) +/- (f_1 +/- f_0).$$

Examination of this formula shows that, by proper selection of upper and lower sideband combinations at J_1, J_2 and J_{02} junction points, we can eliminate the auxiliary frequency f_1 from the combination frequency f_{02} at J_{02} junction. To achieve this, the necessary condition is that the sign, as eventually chosen for the f_1 frequency in (1.1) be opposite to the sign of this frequency in (1.1.1) when the upper sideband at J_{02} junction is selected, or that these signs be the same when the lower sideband at this junction is selected. For instance, if we select $f_{12} = f_2 - f_1$ and $f_{01} = f_1 +/- f_0$, then, for the upper sideband:

$$f_{02} = f_{12} + f_{01} = (f_2 - f_1) + (f_1 +/- f_0) = f_2 +/- f_0. \tag{7}$$

Or if we select $f_{12} = f_2 + f_1$, then, for the lower sideband at J_{02}:

$$f_{02} = f_{12} - f_{01} = (f_2 + f_1) - (f_1 +/- f_0) = f_2 -/+ f_0. \tag{7'}$$

It should be noted that in both cases either of the sideband combinations of f_2 and f_0 frequencies is available.

In a similar manner it can be proved that the frequencies

$$f_{03}, f_{04}, \cdots, f_{0(k-1)}$$

can be obtained as single-sideband combinations of the original low frequency f_0 and of the successive frequencies f_3, f_4, $\cdots$, f_{k-1}, respectively. Eventually, from the next to the last stage $(k - 1)$ we obtain

$$f_{0(k-1)} = f_{(k-1)} +/- f_0.$$

At the last modulation level, from the J_k junction, we obtain

$$f_{(k-1)k} = f_k +/- f_{(k-1)}.$$

By combining these last two frequencies at the J_{0k} junction point, we obtain the final output frequency f_{0k} from the modulation system as

$$f_{0k} = f_{(k-1)k} +/- f_{0(k-1)} = [f_k +/- f_{(k-1)}] +/- [f_{(k-1)} +/- f_0],$$

from which:

$$f_{0k} = f_k +/- f_0 . \tag{8}$$

Thus, the output from the junction J_{0k} will provide the desired single sideband combination of the original frequencies f_k and f_0, with all auxiliary frequency components canceled out.

From the above considerations it is evident that the network shown on Fig. 3 can be extended by an addition of any required number of four-branch meshes. Consequently, the solution it offers has no inherent limitation as to the ratio between two frequencies to be combined. It can be concluded, therefore, that this network represents a general solution to the problem of adding or subtracting two high-ratio ac signal frequencies by means of sequential single-sideband frequency combining.*

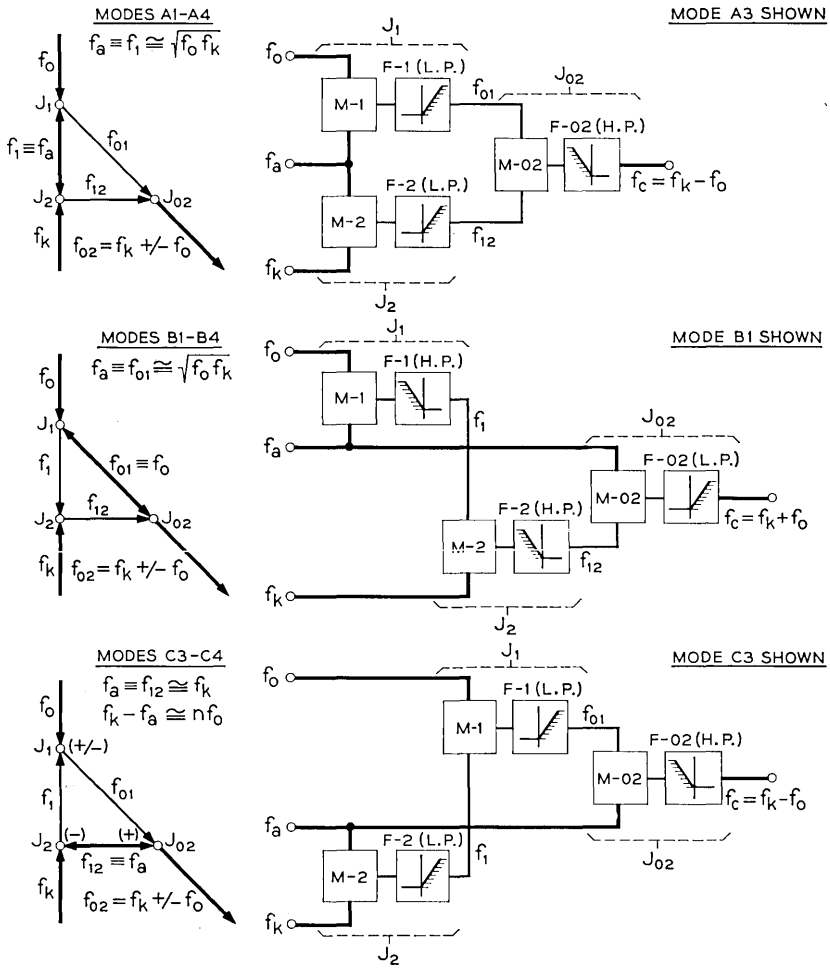
1.3.3 *Possible Network Configurations*

As mentioned above, any branch of a mesh can represent an auxiliary frequency. Thus, by changing the position of the auxiliary frequency within a mesh a certain number of possible network configurations, each of which represents a variant of the general solution, may be obtained. For a given symbolic network, the number of these variants depends upon the number of meshes as given by (4) and thus, eventually, on the parameter k.

It should be noted that the elemental frequency-combining scheme shown in Fig. 1 can be considered as a special case when $k = 1$; as indicated by (4), there is no auxiliary frequency present for this value of k.

For $k = 2$, (4) indicates one mesh, while from (5) the number of junction points can be found as three. In this case, therefore, the symbolic network consists of one three-branch mesh to which the original input frequencies, f_0 and f_k, are applied to two corners and the output frequency is taken from the third corner. Fig. 4 shows three basic network configurations (or *modes*) arrived at when the auxiliary frequency is

* It should be noted that a method of combining multiples or submultiples of two frequencies in order to obtain the sum or difference of these frequencies is excluded from the present considerations. It can be shown, moreover, that frequency "jitters" and phase variations associated with frequency multiplication and division cannot, for inherent reasons, be eliminated from the output frequency.



USEABLE SIDEBAND COMBINATIONS

MODE	J_1	J_2	J_{02}	f_k & f_0 COMBINATION
A1	+	+	−	DIFFERENCE
A2	−	+	−	SUM
A3	−	−	+	DIFFERENCE
A4	+	−	+	SUM
B1	+	+	−	SUM
B2	−	+	−	DIFFERENCE
B3	−	−	+	SUM
B4	+	−	+	DIFFERENCE
C3	−	−	+	DIFFERENCE
C4	+	−	+	SUM

Fig. 4 — Possible network variations.

placed in branches f_1, f_{01} and f_{12} in succession. Additional variations for each of these modes are obtained by choosing different combinations of the sidebands for each of three junction points, as shown in the table in Fig. 4. It could be observed that out of eight (2^3) possible sideband combinations for each mode A, B and C, four combinations must be rejected as not providing cancellation of the auxiliary frequency component at the J_{02} junction point. Thus, the acceptable combinations are only those in which the filter sideband of the J_2 junction point is opposite to the sideband of the J_{02} junction. For configuration C, which uses an auxiliary frequency of the same order as the f_k frequency, two additional combinations must be rejected, since only the lower filter sideband at the J_2 junction can be used to produce an acceptable value for the f_1 frequency. It can be found, therefore, that either the sum or the difference, as desired, of the two original frequencies f_k and f_0, having the ratio index $k = 2$, can be provided in three modes (A, B and C, Fig. 4), two of which (A and B) have two sideband variants each, while the third one (C) has one variant. Thus, the total number of possible circuit arrangements in either case (sum or difference) is five.

Increasing the value of the frequency-ratio index from k to $k + 1$ extends the symbolic network by an additional four-branch mesh. It could be noted that one of the branches of this new mesh is common with the $f_{(k-1)k}$ (see Fig. 3) branch, while the other three branches, f_k, f_{0k} and $f_{k(k-1)}$, are independent. On this basis, the following formula, which expresses the number of modes, M, for any successive $(k + 1)$ order, can be derived:

$$M_k = 4M_{k-1} - M_{k-2}, \tag{9}$$

where M represents the number of basic circuit configuration for a given order.

Thus, for $k = 3$,

$$M_3 = 4M_2 - M_1.$$

But, as stated above, $M_1 = 1$ and $M_2 = 3$. Therefore:

$$M_3 = 4 \cdot 3 - 1 = 11.$$

Similarly, for $k = 4$:

$$M_4 = 4M_3 - M_2 = 4 \cdot 11 - 3 = 41.$$

Thus, by using (9), the total number of possible circuit configurations can be found for any symbolic network of the $(k + 1)$ order.

II. APPLICATIONS

2.1 *Design Considerations*

2.1.1 *Practical Circuit Configurations*

A relatively large number of possible circuit arrangements, all of which are derivable from the symbolic frequency-combining diagram, may present considerable practical advantages. Since these arrangements can be systematically scrutinized, in any given case the most practical solution can be thus attained. Such a solution may take into account certain particular requirements — for instance, low content of modulation products, avoidance of phase and amplitude distortions and, especially, availability of certain components such as filters and oscillators.

In the design of systems concerned with frequency selection and discrimination, the problem of filter characteristics is usually one of the important considerations. In general, electric wave filters of a special design are relatively expensive, particularly when stringent requirements must be met. The method presented here allows the use of filters of any limited frequency discrimination characteristics in building systems that realize much higher frequency discrimination capabilities. Thus, the necessity for using filter networks of a difficult or impractical design is eliminated. Eventually, an optimal solution compromising the quality of the filters to be used in the given system with their over-all number can be established. Moreover, in many cases filters of an available design may be utilized.

Adaptability of this procedure to certain practical cases is illustrated in the examples of Section 2.2 below.

2.1.2 *Effects of Excessive Auxiliary Frequency Drifts*

As explained above, drifts of the auxiliary frequencies do not essentially affect stability of the output signal of the desired frequency combination. However, since variations of the auxiliary frequencies do affect the positions of the operational points within the filter passbands, excessive drifts may, under certain practical conditions, produce objectionable secondary effects in the form of phase and amplitude distortions. This may take place when the incremental characteristics of attenuation and phase versus frequency of certain pairs of filters carrying the same auxiliary frequencies differ appreciably within the bandwidth covered by the shifts of the operational points.

These amplitude and phase distortions, which rarely exceed acceptable limits, can be reduced, if necessary, in several ways such as:

(a) stabilization to a desirable degree of the auxiliary frequency sources;

(b) shifting, by a discrete amount, of any particular auxiliary frequency value so that an optimal operational section of the filter characteristics can be used;

(c) corrective shaping of the filter passband characteristics so that compensating effects can be achieved;

(d) introducing amplitude and phase equalizers at the auxiliary oscillator outputs.

By applying the above methods, separately or in combination, phase and amplitude distortions usually can be reduced to acceptable levels.

2.1.3 *Noise Levels*

It has been found experimentally that, by observing precautions usually applied in the design of high-quality transmission circuits, noise levels at the output from the systems described here are particularly low. This may be attributed to the fact that the virtual bandwidth of these systems actually is very narrow as a result of certain sequential combinations of filter networks. In many practical cases this property can be enhanced still further by using bandpass filters rather than low- or high-pass filters, as called for by the essential design requirements. Examples of such modifications, which are also desirable when certain filters of a standard design are available, are shown below.

2.2 *Practical Examples*

The following examples fall into three categories, which embrace some typical problems encountered in practice:

(a) combining two essentially fixed high-ratio signal frequencies;

(b) combining two high-ratio signal frequencies, one of which is adjustable over a wide frequency range;

(c) providing high-accuracy signal frequency sources that can be varied over wide frequency ranges either in digital or arbitrarily chosen frequency steps.

2.2.1 *Combining Two High-Ratio Essentially Fixed Frequencies*

In an application related to the phase-delay measurements, a solution of the following problem is required (see Fig. 5):

A frequency of 90 mc with drifts and random frequency variations of

the order of ±1000 cps represents a "master" frequency of a measuring system. It is expected that some of these variations may occur at a rate too rapid for any mechanoelectronic servo system to follow without excessive tracking errors. It is required that another ac signal source be provided, the frequency of which should be higher than the master oscillator frequency by an increment of 55,560 ±10 cps. Thus, using denota-

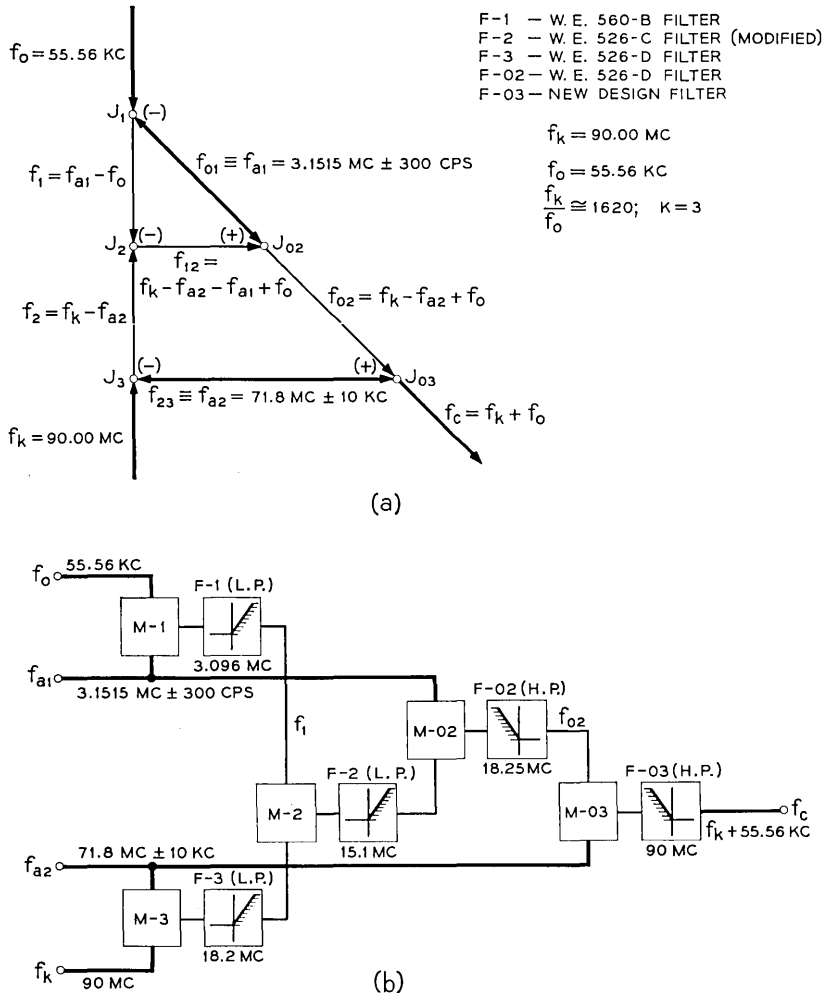


Fig. 5 — Fixed-range "slave" oscillator system: (a) symbolic diagram; (b) block diagram.

tions previously introduced:

$$f_k = 90{,}000 \text{ kc},$$

$$f_0 = f_c - f_k = 55.56 \text{ kc},$$

$$(n')^k = \frac{f_k}{f_0} \cong 1620.$$

For the solution of this problem, the following filters of an available design (manufactured by the Western Electric Company) were taken under consideration:

(a) Type 560-B filter with a passband of 3.096 mc $\pm$1000 cps; n value over 100 (for 60-db minimum discrimination).

(b) Type 526-C filter (modified) with a passband at 15.1 mc $\pm$100 kc; n value of the order of 10.

(c) Type 526-D filter with a passband at 18.2 mc $\pm$100 kc; n value of the order of 10.

Using these filters, the k value has been found as follows:
Assuming

$$n_1 \cong 100,$$

$$n_2 \cong 10,$$

$$n_3 \cong 10,$$

$$n_1 n_2 n_3 = 100 \cdot 10 \cdot 10 > 1620.$$

Then,

$$k = 3,$$

and thus the number of auxiliary frequencies, from (4), is

$$m_a = k - 1 = 2$$

and the number of junction-points, from (5), is

$$m_j = 2k - 1 = 5.$$

Subsequently, the symbolic network shown in Fig. 5 was designed. From a study of this network it was found that an optimal use of the available filters may be achieved by placing one 3.096-mc filter (560-B) as a low-pass filter at the J_1 junction point, an 18.2-mc filter (526-D) as a low-pass filter at the J_3 point, and another filter of the same type as an essentially high-pass filter at the J_{02} point. Finally, one 15.1-mc filter (modified 526-C) was placed at the J_2 point as a low-pass filter.

By locating one of the auxiliary frequencies (3.1515 mc ±300 cps) in the f_{01} branch and the other (71.8 mc ±10 kc) in the f_{23} branch, all the other frequencies were found according to the rules described above. The only filter for which no existing design had been found is at the J_{03} junction-point. This filter, however, must pass the signal of 90.00 mc as the sum of f_{02} frequency (approximately 18.2 mc) and f_{23} frequency (minimum 71.8 mc), while rejecting the other modulation products. As these requirements indicate, the design of such a filter network is well within practical limits of art.

Fig. 5(b) shows a conventional block diagram, representing the practical solution derived as described, from the symbolic network of Fig. 5(a).

2.2.2 *Combining Two High-Ratio Signal Frequencies, One of Which is Adjustable Over a Discrete Range*

An example of a solution applicable to this case is shown in Fig. 6. The frequency of a "master" ac signal source, covering the range from 20 to 100 mc, has to be followed by a "slave" ac signal source having a frequency higher than the "master" frequency by a constant amount of 55.556 kcs.

Let

$$f_{v\,\max} = 100 \text{ mc},$$

$$f_{v\,\min} = 20 \text{ mc},$$

$$f_0 = 55.556 \text{ kc}.$$

The solution in this case can be arrived at in two steps. In the first step, a fixed frequency, f_k, is established so that a single-stage combining of this frequency with the frequency difference, $f_k - f_v$, is realizable in accordance with (2). Thus,

$$\frac{f_k}{f_k - f_{v\,\max}} = n_k',$$

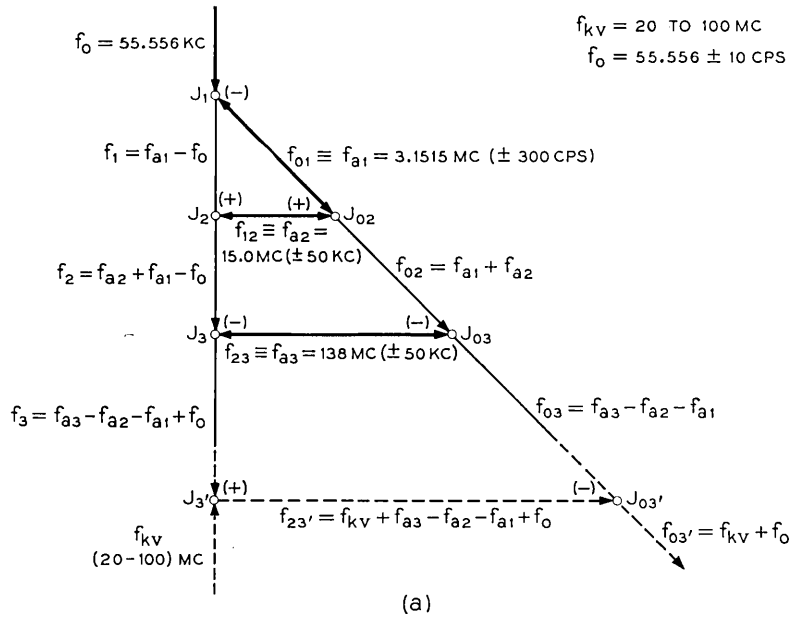
which gives

$$f_k = \frac{n_k'}{n_k' - 1} f_{v\,\max}. \tag{10}$$

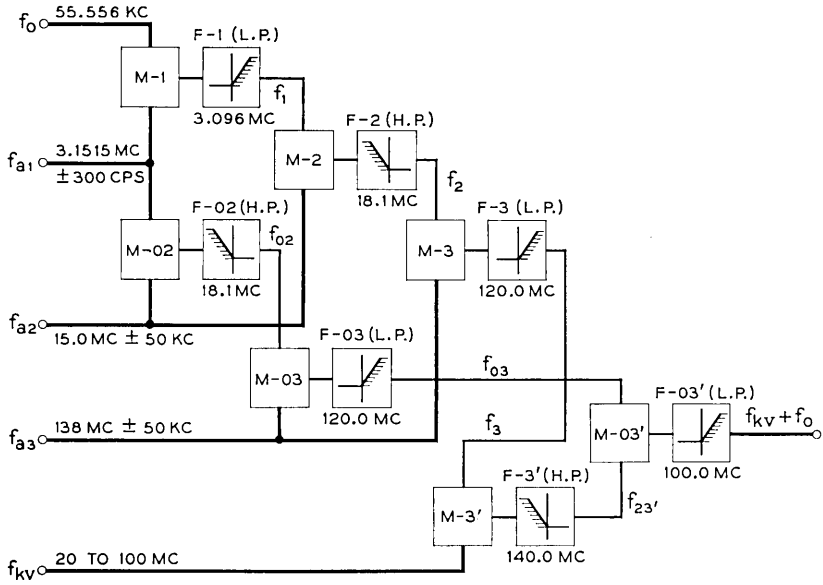
On the other hand, since the whole range of the variable frequency f_v, down to its minimum value $f_{v\,\min}$ must be combined with the f_k frequency, then:

$$\frac{f_k}{f_{v\,\min}} \leqq n_k',$$

or



(a)



(b)

Fig. 6 — Variable-range "slave" oscillator system: (a) symbolic diagram; (b) block diagram.

$$f_k \leqq n_k' f_{v\,\min}\,. \tag{11}$$

Assuming $n_k' \cong 6$, f_k can be found to be 120 mc.

In the second step the problem is thus reduced to adding two fixed signal frequencies of 55.556 kc (f_0) and of 120 mc (f_k) to each other. But this can be done in a manner similar to that described in the previous example. Thus,

$$\frac{f_k}{f_0} = \frac{120{,}000}{55.556} \cong 2200\,.$$

Assuming, as previously, that $n_1 \cong 100$ and $n_2 \cong n_3 \cong 10$, we obtain

$$n_1 n_2 n_3 = 100 \cdot 10 \cdot 10 \geqq 2200.$$

Then

$$k = 3\,.$$

Subsequently, the symbolic diagram, as shown by the solid lines in Fig. 6(a), can be constructed. To this diagram, a mesh comprising the J_3' and J_{03}' junction points is added, as shown by dotted lines. The variable "master" signal frequency is connected to the J_3' junction point (which comprises a filter with discrimination ratio of n_k'); and the variable "slave" signal frequency is available from the J_{03}' junction point (which comprises a similar filter).

One of the possible practical versions of this solution that could be derived from the symbolic network of Fig. 6(a), is shown in Fig. 6(b).

2.2.3 *Continuously Adjustable Signal Frequency Sources*

This special case of circuitry, which can be derived simply from a single-mesh frequency-combining network (see Fig. 4, mode C), is shown in Fig. 7. It provides a continuously adjustable range of signal frequencies, these being the sum or the difference of a desired harmonic of a standard frequency signal and an interpolating frequency source. In place of the signal frequency f_0, a low-frequency signal source that is continuously variable over a frequency interval equal to a fundamental frequency value is connected. In place of the signal frequency f_k, a multifrequency signal source in the form of a harmonic generator that supplies harmonics of the same fundamental frequency value is connected. Subsequently, in place of the f_{12} mesh frequency, a special auxiliary frequency source is applied. The frequency of this source can be varied in steps, each being essentially equal to the same fundamental frequency value. At the J_2 junction point, a narrow passband filter is placed. The center point of its passband is equal to the sum of one of

the harmonics present at f_k and an auxiliary frequency available at f_{12}. In this manner, the following results are achieved:

i. The f_1 frequency derived from the combination of one of the harmonics and an auxiliary oscillator frequency remains essentially constant while successive harmonics enter the combination.

ii. In order to obtain f_{01} and f_{02} as single sideband frequencies, only fixed low-pass or high-pass filters must be used.

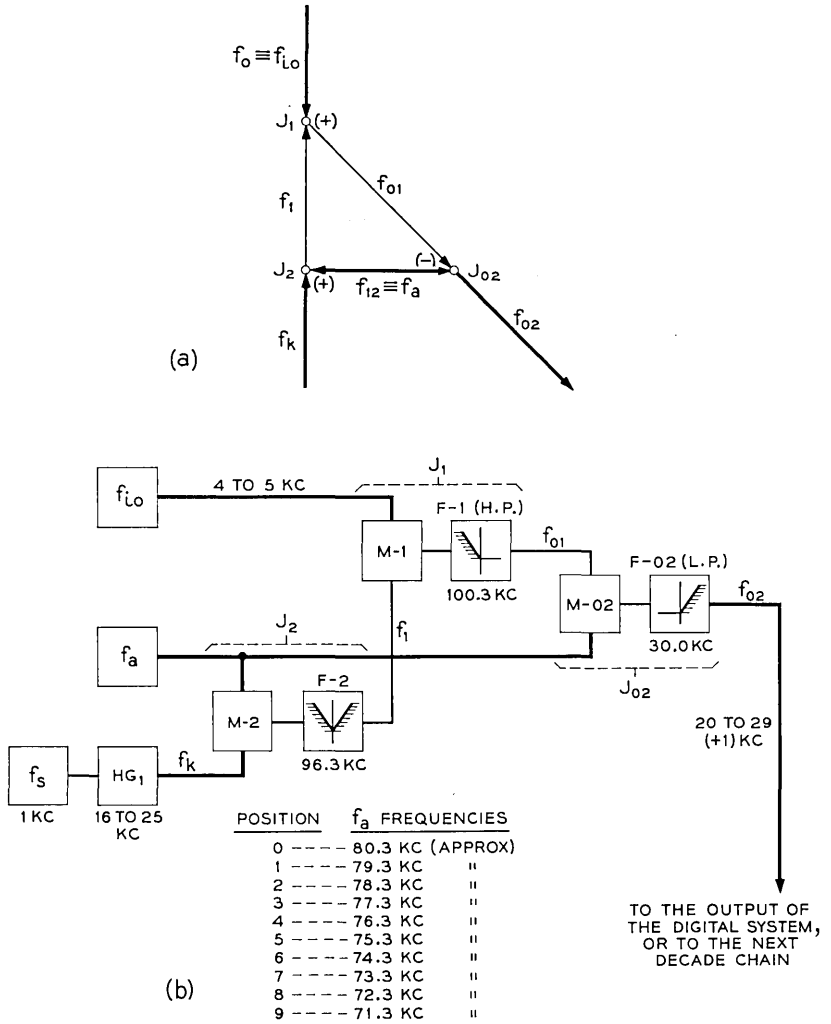


Fig. 7 — Decade "chain" (digital frequency-combining and frequency-selecting system: (a) symbolic diagram; (b) block diagram.

iii. At the output from the J_{02} junction point, a signal of a single sideband frequency derived as the sum or difference of a desired harmonic frequency and the interpolating oscillator frequency is available.

The above properties of this circuitry allow the design of precise signal frequency systems which cover extensive frequency ranges with any desired degree of resolution. In contrast to other systems,[10] the frequency ranges are adjustable here in digital steps without using tunable elements.

An example of an elemental practical circuit[11] is shown in Fig. 7(b). It represents a "step-up digital chain" covering the frequency range from 20 to 30 kc by 10-decade steps and an interpolating oscillator.

The interpolating oscillator covers the range from 4.000 to 5.000 kc and is calibrated in such a way that the "0" cps mark corresponds to 4.000 kc and the "1000" cps mark corresponds to 5.000 kc. A harmonic generator, connected to a 1-kc high-accuracy frequency standard, provides harmonics utilized in the range from the 16th through the 25th. The auxiliary frequency is controlled by the 0–9 positions of a decade switch and provides 10 frequencies, from 80.3 to 71.3 kc respectively, in approximately 1-kc intervals.

The proper operation of this step-up chain depends primarily on the filter at J_2 junction point (F-2 filter). It is a narrow passband filter that rejects the fundamental and all harmonics of the frequency standard available from the harmonic generator, but which passes a narrow frequency band between the numerical values of two successive integral multiples of the standard frequency. It is desirable that this passband be located well beyond the useful range of the harmonic generator, possibly in the vicinity of the point where the envelope of the harmonic spectrum is near, or crosses, the zero amplitude value. The types of filters particularly suited to this application are some of the quartz crystal filters developed for the carrier telephony[12] or mechanical filters used for certain communication radio receivers of an advanced type.[13] The filter used in a practical application of this case is the Western Electric Company's standard crystal filter (97A), which has the passband frequency at 96.3 kc and a flat characteristic within ± 50 cps.

As an auxiliary frequency source, an oscillator adjustable in 10 steps of approximately 1-kc each is used. Its frequency, corresponding to the "0" decade step, is thus 80.3 kc, while the frequency corresponding to the ninth decade step is equal to 71.3 kc.

The F-1 and the F-02 filters are somewhat modified versions of standard carrier telephony filters.

The operation of the circuit, as shown in Fig. 7(b), may be illustrated

by a numerical example. Let us assume that the setting of the frequency of 26.785 kc is desired. Thus, the 1-kc decade dial is set to position 6, in which setting a signal of 74.3 kc $\pm d_1$ (where d_1 represents an instantaneous value of the drift of the auxiliary frequency) is provided. The interpolating oscillator, set to "785" cps, provides a 4.785-kc signal. Subsequently, the F-2 filter passes the 96.3-kc $\pm d_1$ signal, which in this case will be a combination of the 74.3-kc $\pm d_1$ signal from the auxiliary oscillator and the 22nd harmonic of the 1-kc standard frequency signal. The adjacent signals of 95.3 kc $\pm d$ and 97.3 kc ± 1, as well as any other signals formed by the auxiliary frequency and the 1-kc harmonic spectrum available from the harmonic generator, are suppressed by the filter F-2 with substantially higher than 60-db attenuation.

In the modulator M-1 the signal of 96.3 kc $\pm d_1$ is combined with the 4.785-kc signal from the interpolating oscillator. The difference of these two signals, equal to 91.515 kc $\pm d_1$, is suppressed by the high-pass filter F_1 having cutoff frequency of 100.3 kc, while their sum, equal to 101.085 kc, passes this filter and is applied to the modulator M-02. In this modulator the same auxiliary frequency of 74.3 kc $\pm d_1$ is combined with the 101.035-kc $\pm d_1$ signal from the filter F-1. The sum of these two signals, equal to 175.385 kc $\pm 2d_1$, is rejected by the low-pass filter F-02 having a cutoff frequency of 30.0 kc. The difference of these two signals,

$$101.085 \pm d_1 - (74.3 \pm d_1) = 26.785 \pm 0,$$

is thus available at the filter output. It may be noted that the cancellation of the auxiliary frequency drift took place as the result of adding and then subtracting the auxiliary frequency, and that the desired output signal frequency is eventually the sum of the 22nd harmonic of the 1-kc standard frequency and the interpolating oscillator output only.

A number of step-up frequency digital chains, like the one just described, can be designed for various frequency levels that are related to each other by a desired order of any digital system. Each chain of the lower order can be connected as a source of the interpolating frequency for the higher chain. By combining a necessary number of such chains, wide-range systems having digital frequency readability and resetability and any desired degree of resolution can be thus realized. An example of a 3-decade system is shown in Fig. 8.

The decade chains of multidecade systems are similar to the one described above, with the exception that some of the auxiliary oscillator decade steps of the higher decades may be suppressed. Also, the auxiliary oscillators of the intermediate decades provide two alternative signal frequencies, depending upon the decade control settings of the higher

decades. Detailed explanation and theory of these decade arrangements can be found in Ref. 11.

Output switching from the individual decade chains may also be coupled to the decade controls so that eventually a continuous coverage of the predetermined frequency range at the output terminals of the system may be achieved.

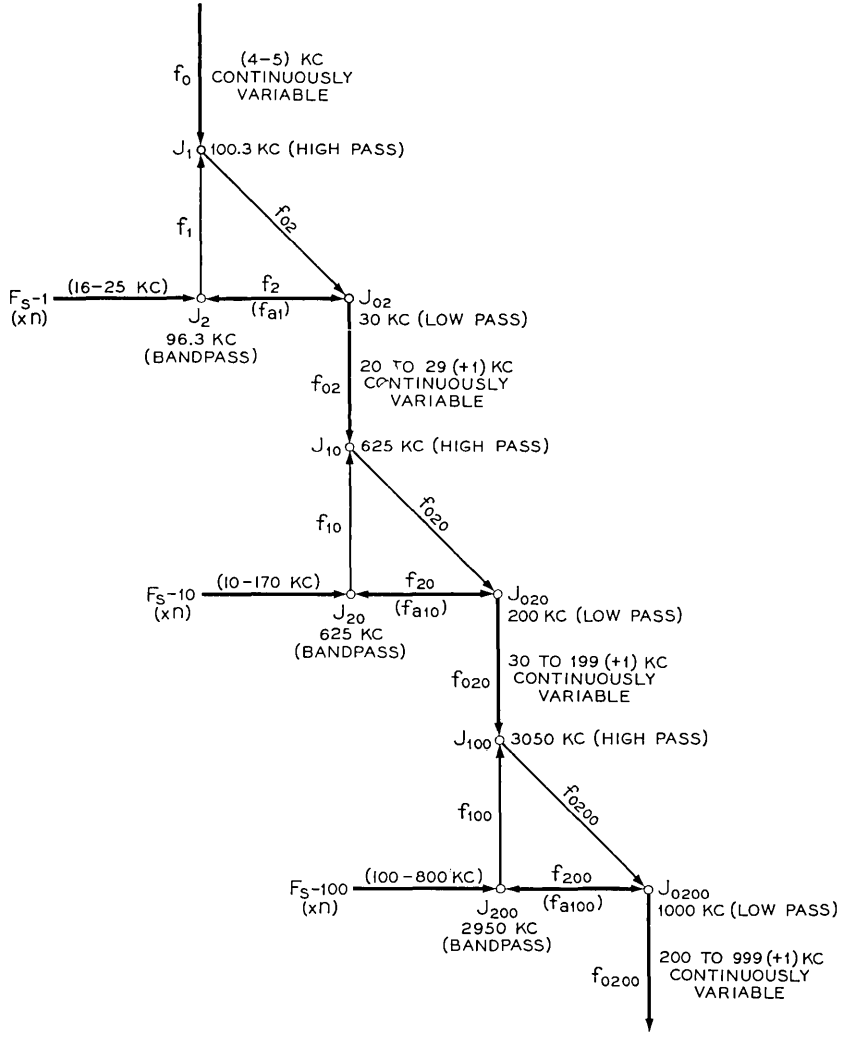


Fig. 8 — Symbolic diagram of a multidecade frequency-combining system.

III. CONCLUSIONS

The theory of a frequency-combining network presented here leads to a general solution of the problem of adding or subtracting two signal frequencies, independently of their ratio, while using electrical wave filters of a practical design with arbitrarily limited frequency discrimination characteristics.

A special symbolic network, established by means of a "sideband algebra", is applicable to signal frequency-combining problems. From this network not only can the minimum number of basic circuit elements be found, but also the fundamental circuitry and all its possible practical variations can be derived and scrutinized.

IV. ACKNOWLEDGMENTS

The author wishes to express his appreciation to G. F. Critchlow, T. H. Crowley, S. Doba, J. O. Israel and S. J. Zammataro for encouragement and helpful comments given to him during the work on this paper.

REFERENCES

1. Lewis, F. D., Frequency and Time Standards, Proc. I.R.E., **43,** September 1955, p. 1046.
2. Peterson, E. and Llewellyn, F. B., The Operation of Modulators from a Physical Viewpoint, Proc. I.R.E., **18,** January 1930, p. 38.
3. Bode, H. W. and Dietzold, R. L., Ideal Wave Filters, B.S.T.J., **14,** April 1935, p. 215.
4. Sterky, H., Frequency Multiplication and Division, Proc. I.R.E., **25,** September 1937, p. 1153.
5. Favre, R., Wideband Frequency Divider, Helv. Acta Physica, **28,** May 31, 1955, p. 172.
6. Meacham, L. A., High-Precision Frequency Comparisons, The Bridge of Eta Kappa Nu, **36,** February-March 1940, p. 5.
7. Shaull, J. M., High-Precision Automatic Frequency Comparator and Recorder, Tele-Tech & Elect. Ind., **14,** January 1955, p. 58.
8. Alsberg, D. A. and Leed, D., A Precise Direct Reading Phase and Transmission Measuring System for Video Frequencies, B.S.T.J., **28,** April 1949, p. 221.
9. Sternberg, R. L. and Kaufman, H., A General Solution of the Two-Frequency Modulation Product Problem, J. Math. & Phys., **32,** January 1954, p. 233.
10. Finder, H. J., Frequency Generation and Measurement, Elect. Engineer, **22,** June 1950, p. 220.
11. Wojciechowski, B. M., U.S. Patent No. 2,745,962.
12. Lane, C. E., Crystal Channel Filters for the Cable Carrier System, B.S.T.J., **17,** January 1938, p. 125.
13. Roberts, W. V. and Burns, L. L., Jr., Mechanical Filters for Radio Frequencies, RCA Rev., **10,** September 1949, p. 348.

An Evaluation of AM Data System Performance by Computer Simulation

By R. A. GIBBY

(Manuscript received February 9, 1960)

The mathematical relationships that describe an amplitude-modulated data system are developed in a form suitable for programming on a high-speed digital computer. These equations contain expressions that specify in general terms the transmission-frequency characteristics of a transmission medium. A data signal composed of a train of raised-cosine shaped pulses is generated in the stimulating process. The simultation provides a means for computing the resulting response of systems to pulse trains. The performance of a double-sideband AM data system is evaluated from measurements of the maximum vertical opening, or aperture, of the eye pattern formed by the received signal. This aperture is related to the system performance in terms of signal-to-noise ratio and error rate of the system. A verification of this technique is made by simultaing the conditions of an experimental laboratory data system on the computer and comparing computed and measured performance.

I. INTRODUCTION

The process of transmitting digital information and the problems relating to this type of communication have received much attention in recent years. Practical means of communication based on this type of information transmittal have been in use for a long time in the form of telegraph and teletype systems, but for both these systems the speed of operation is relatively slow, since the information is supplied either by a manual operator or by a mechanical device of some sort. Recently, however, many new sources of digital information have arisen for which the information rates may range from telegraph speed up to several orders of magnitude greater than that speed. This has created a need for new high-speed data transmission systems. To this need for higher speed has

been added the requirement of greater accuracy — amounting to, in some cases, essentially error-free transmission. For a given bandwidth both high speeds and accurate transmission are often difficult to realize. For this reason, a great deal of effort has been spent in devising means for utilizing a channel as effectively as possible in order that both speed and accuracy may be maximized.

This effort may be divided into two areas. The first is concerned with what may be called the terminal problem, in which effort is directed toward developing data system terminal equipment that makes best use of the transmission channel. This has resulted in a number of competing schemes for data transmission based on different modulation methods, such as double and vestigial sideband amplitude modulation, frequency modulation and phase modulation. Also, a number of different detection methods are available for these various systems.

The second area of effort may be referred to as the transmission problem. Here, work is directed toward determining what influence the various transmission factors such as the transmission-frequency characteristics and noise have on the transmitted data signal, and what conditions should be maintained for a particular data system that is to operate at a given speed and accuracy.

In order to answer questions that arise in these two areas, it is necessary that a method for evaluating data system performance be devised. Good performance is associated with both high speed and accurate transmission, but quality of performance is not necessarily preserved by a direct interchange of these two quantities. It is of little value to send information at very high speeds with very low accuracy. The optimum performance is obtained, therefore, by transmitting data over a given channel at the highest rate possible, consistent with a given error criterion as dictated by the system requirements.

Because of the many factors that contribute to the over-all performance, the problem of evaluating performance is a difficult one. The most direct method and the one that has been most widely used is that of measuring performance on a data system under actual operating conditions. Some results from such experimental tests have been given[1] and made general, at least to some extent. The experimental results, however, are of necessity limited in their generality, and the task of obtaining results by this method is expensive and time-consuming. In addition, for design purposes it is desirable to know how a data system will perform before it is built as well as after.

Another approach to evaluating performance is by means of a theoretical analysis of the problems involved. A considerable amount of work

has been done in this area.[2,3,4,5] While this work is of great value, the complexity of the problems make a general solution untractable, except under conditions of rather restricted and idealized assumptions. For conditions of practical significance, however, a system analysis using mathematical models is useful, provided the task of a numerical computation can be carried out.

This paper is the result of an effort to evaluate the performance of an AM data system by methods of numerical analysis and simulation of the system on a digital computer. In this work, the general method of analysis and simulation is outlined, a criterion for performance is defined and some results that have been obtained by the use of this method are given. While the problem is analyzed in rather general terms, the particular results presented apply to the specific problem of evaluating the degradation in performance due to delay distortion.

II. BASIC DATA SYSTEM

The basic data system considered here is shown in Fig. 1. The input function $f(t)$ is the electrical representation of the random sequence of binary information to be transmitted. This signal is made up of a succession of identical raised-cosine shaped pulses interspersed with gaps or spaces where no pulse occurs, forming a pattern of the type illustrated in Fig. 2. In this pattern the occurrence or absence of a pulse is governed by the occurrence of a one or a zero in the binary information. These

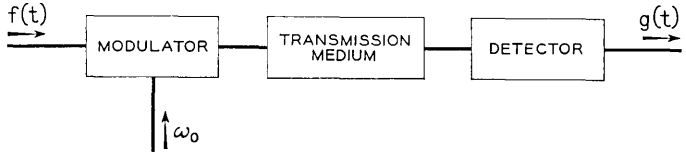


Fig. 1 – Data transmission system.

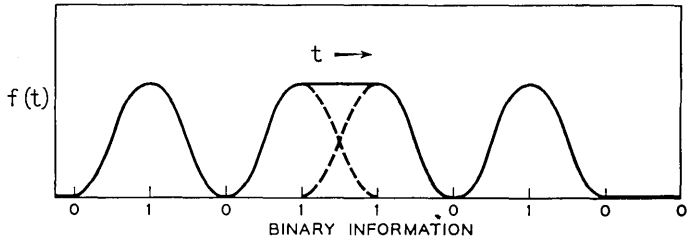


Fig. 2 – Binary information and equivalent pulse pattern.

binary values correspond in time to the points at which pulses may occur in the pattern, and they are separated by a uniform time interval called the *sampling interval*, which is the reciprocal of the bit speed of the system.

The basic pulse used to form the pattern is shown in Fig. 3. It is chosen to have a maximum width t_0, which is equal to two bit intervals at its base. Because of this fact, and because of the raised-cosine shape, pulses that are adjacent in the pattern of Fig. 2 add together over the interval that separates them to give a constant value equal to the maximum value of a single pulse. Also, adjacent gaps produce a constant value equal to zero. For the case of alternating values of one and zero, transitions are produced that are portions of a sinusoid.

For purposes of analysis it will be useful to consider a single pulse of the type used in forming the pulse pattern of Fig. 2. To do this a particular pattern is used, as shown in Fig. 3(a), in which $f(t)$ is formed by the binary sequence 010 followed by a sequence of zeros and then repeated periodically at intervals of T. This method of specifying the signal permits the analysis of the basic pulse to be made on a Fourier series basis that is conveniently adapted to numerical computation on the digital computer.

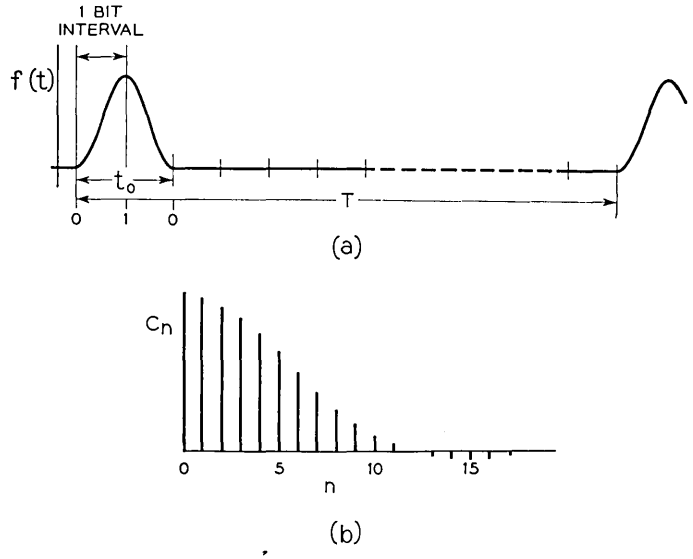


Fig. 3 – Basic pulse used to form pattern: (a) raised-cosine pulse; (b) associated spectrum.

The spectral density function resulting from a Fourier analysis of the periodic pulse pattern of Fig. 3(a), is shown in Fig. 3(b), and is expressed mathematically as

$$C_n = \frac{\sin \frac{n\pi t_0}{T}}{n\pi \left[1 - \left(\frac{nt_0}{T}\right)^2\right]}. \tag{1}$$

We can imagine a perfect data system in which the signal $f(t)$ of Fig. 2 is transmitted without alteration. This will be considered a reference condition for measuring performance. Under this condition the binary information can be extracted from the received signal $g(t)$, which is identical to $f(t)$, by examining the pulse pattern at the sampling times to determine if the signal is greater than or less than some intermediate-value called the *slice level.* (In most cases this level is equal to one half the value of the maximum pulse height. Under certain conditions, however, it will be advantageous to use other values.)

For a system with ideal transmission-frequency characteristics, errors are produced if a sample is made during the time when a noise disturbance is large enough to cause the signal to make an erroneous excursion across the slice level. In order to meet system requirements the value of the signal-to-noise ratio must be such that the number of errors in a given time does not exceed a certain maximum. If the signal-to-noise ratio is less than this value the system is considered as failing. If the signal-to-noise ratio is such that the error criterion is just met, then a reduction in the noise (with the signal held constant) results in an operating margin against failure equal to the amount of the noise reduction. If the system were operating without noise, the amount of noise necessary to bring it to the failing condition would be called the *noise margin* for the system.

As the transmission-frequency conditions depart from ideal, the output $g(t)$ will no longer be identical to the input $f(t)$. This lack of fidelity is evidence of system degradation. It will be shown later how this degradation can be expressed quantitatively as a reduction in the noise margin for the system.

Transmission of the data signal over most facilities requires some form of modulation. It will be assumed here that the signal is amplitude-modulated by an ideal product modulator, and that an ideal envelope detector is used at the receiver. Also, the system will be considered, for the present, as being noise-free, and the impairment that the data signal suffers in transmission will be due entirely to the frequency characteristics of the transmission medium.

III. TRANSMISSION CHARACTERISTICS

The basic relationship that describes the transmission-frequency characteristics of a transmission system is given as

$$T(\omega) = A(\omega)e^{-j\psi(\omega)}, \tag{2}$$

in which $A(\omega)$ is the attenuation and $\psi(\omega)$ the phase characteristic. In the work to follow we will make use of the fact that the attenuation characteristic has even symmetry $A(\omega) = A(-\omega)$ and the phase characteristic has odd symmetry $\psi(\omega) = -\psi(-\omega)$ about the zero frequency axis.

To accommodate the use of a Fourier series development of the problem, the transmission-frequency characteristics can be expressed in terms of their values at discrete radian frequencies corresponding to the Fourier series components, rather than on a continuous basis. For this purpose, (2) becomes

$$T\left(\frac{n2\pi}{T}\right) = A\left(\frac{n2\pi}{T}\right)e^{-j\psi(n2\pi/T)}. \tag{3}$$

The phase curve of Fig. 4(a) has a shape typical of the characteristics of many transmission facilities and can be approximated by a sine-wave departure from linearity as given by

$$\psi(\omega) = a\omega - b \sin(\omega\tau + \theta), \tag{4}$$

where:

b is the maximum departure of the phase from linearity,

τ is the reciprocal of the period of the phase curve and

θ is the displacement of the phase curve from symmetry about the carrier.

The corresponding envelope delay is shown in Fig. 4(b) and is expressed analytically as the derivative of the phase:

$$\frac{d\psi}{d\omega} = a - b\tau \cos(\omega\tau + \theta). \tag{5}$$

The linear term in the phase expression results in only a constant delay, and the distortion in transmission is therefore produced entirely by the sine-wave departure from linearity. The three parameters b, τ and θ give a complete description of any sinusoidal delay or phase characteristic. A wide variety of such curves can be realized by letting these three parameters range over values of interest (it will be shown later that, with a sinusoidal representation of phase, a very close approximation to actual transmission characteristics can be achieved). With this repre-

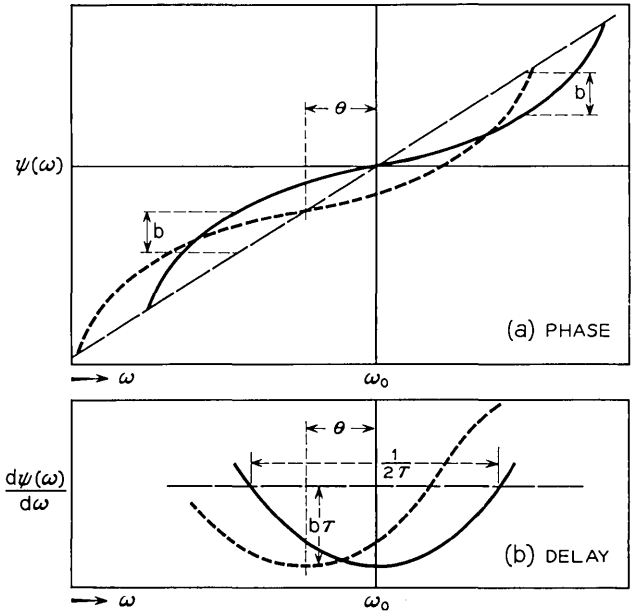


Fig. 4 – Sinusoidal characteristics: (a) phase; (b) delay.

sentation of the transmission characteristics, the problem of determining the performance of a data system in terms of the parameters b, τ and θ will now be considered.

IV. ANALYSIS OF AN AM DATA SYSTEM

In this section,* the analysis of an AM system is made in order to obtain the expression for the output function $g(t)$ in terms of the input function $f(t)$ and the parameters b, τ and θ. For a periodic input pulse as shown in Fig. 3(a) the output response will also be periodic. If the period T is chosen large enough to allow the transient response of the output to be restored essentially to zero there will be no intersymbol interference between these periodic pulses. We will then have a condition in which both $f(t)$ and $g(t)$ can be represented by a Fourier series.

The signal that results from an amplitude modulation of $f(t)$ by a cosine carrier is given by

$$h(t) = f(t) \cos \omega_0 t. \tag{6}$$

* This analysis is based on an unpublished memorandum of R. G. Segers. It includes some slight modification for use in the present study.

Expressing $f(t)$ by a complex Fourier series representation and using the complex form of the cosine function, we have

$$h(t) = \frac{e^{j\omega_0 t} + e^{-j\omega_0 t}}{2} \sum_{n=-\infty}^{\infty} C_n e^{jn(2\pi/T)t}, \tag{7}$$

where

$$C_n = \frac{1}{T} \int_{-T/2}^{T/2} f(t) e^{-jn(2\pi/T)t}\, dt. \tag{8}$$

Equation (7) may be expressed as

$$h(t) = \frac{1}{2} \sum_{n=-\infty}^{\infty} C_n e^{j[n(2\pi/T)+\omega_0]t} + \frac{1}{2} \sum_{n=-\infty}^{\infty} C_n e^{j[n(2\pi/T)-\omega_0]t}. \tag{9}$$

The modulated signal expressed by (9) is composed of two doubly infinite spectra, one distributed about the positive carrier of radian frequency ω_0, and the other about the negative carrier of radian frequency $-\omega_0$. It is apparent that this representation is redundant, and for purposes of efficient computation the number of terms can be reduced by a factor of two. This will be done in a way that preserves certain desired symmetries in the equations and allows the results to be interpreted meaningfully. First, however, it is necessary to determine the carrier response by introducing into (9) the expression for the transmission-frequency characteristics. The transmission-frequency characteristics given by (3) relate to baseband conditions; in order to properly relate this expression to carrier transmission it is necessary to translate these characteristics by an amount of $+\omega_0$ and $-\omega_0$. Making this transformation in (3) and substituting the results into (9) gives the following:

$$\begin{aligned} g(t) = {} & \frac{1}{2} \sum_{n=-\infty}^{\infty} C_n A \left[n\left(\frac{2\pi}{T}\right) + \omega_0 \right] e^{j[n(2\pi/T)+\omega_0]t} e^{-j\psi[n(2\pi/T)+\omega_0]} \\ & + \frac{1}{2} \sum_{n=-\infty}^{\infty} C_n A \left[n\left(\frac{2\pi}{T}\right) - \omega_0 \right] e^{j[n(2\pi/T)-\omega_0]} e^{-j\psi[n(2\pi/T)-\omega_0]}, \end{aligned} \tag{10}$$

which is the expression of the response of the carrier signal to the transmission system.

Use is now made of the fact that $f(t)$ is an even function. The quantity C_n, which can, therefore, be expressed by taking twice its value over one half of its range, is given as,

$$C_n = \frac{2}{T} \int_0^{T/2} f(t) \cos n\left(\frac{2\pi}{T}\right) t\, dt. \tag{11}$$

When this, and also the fact that the attenuation must have even sym-

metry and the phase must have odd symmetry, are taken into account, (10) can be rewritten for $n \gtrless 0$, giving

$$\begin{aligned} r(t) = &\sum_{n=0}^{\infty} C_n A\left[n\left(\frac{2\pi}{T}\right)+\omega_0\right]\cos\left\{\left[n\left(\frac{2\pi}{T}\right)+\omega_0\right]t \right. \\ &\qquad \left. -\psi\left[n\left(\frac{2\pi}{T}\right)+\omega_0\right]\right\} \\ &+\sum_{n=1}^{\infty} C_n A\left[n\left(\frac{2\pi}{T}\right)-\omega_0\right]\cos\left\{\left[n\left(\frac{2\pi}{T}\right)-\omega_0\right]t \right. \\ &\qquad \left. -\psi\left[n\left(\frac{2\pi}{T}\right)-\omega_0\right]\right\}. \end{aligned} \tag{12}$$

Rather than computing with (12) as it stands, it will be more useful to extract the high-frequency carrier by the use of elementary trigonometric relationships, giving

$$\begin{aligned} r(t) = &\cos\omega_0 t\sum_{n=0}^{\infty} C_n A\left[n\left(\frac{2\pi}{T}\right)+\omega_0\right] \\ &\qquad \cdot\cos\left\{n\left(\frac{2\pi}{T}\right)t-\psi\left[n\left(\frac{2\pi}{T}\right)+\omega_0\right]\right\} \\ &-\sin\omega_0 t\sum_{n=0}^{\infty} C_n A\left[n\left(\frac{2\pi}{T}\right)+\omega_0\right] \\ &\qquad \cdot\sin\left\{n\left(\frac{2\pi}{T}\right)t-\psi\left[n\left(\frac{2\pi}{T}\right)+\omega_0\right]\right\} \\ &+\cos\omega_0 t\sum_{n=1}^{\infty} C_n A\left[n\left(\frac{2\pi}{T}\right)-\omega_0\right] \\ &\qquad \cdot\cos\left\{n\left(\frac{2\pi}{T}\right)t-\psi\left[n\left(\frac{2\pi}{T}\right)-\omega_0\right]\right\} \\ &+\sin\omega_0 t\sum_{n=1}^{\infty} C_n A\left[n\left(\frac{2\pi}{T}\right)-\omega_0\right] \\ &\qquad \cdot\sin\left\{n\left(\frac{2\pi}{T}\right)t-\psi\left[n\left(\frac{2\pi}{T}\right)-\omega_0\right]\right\}. \end{aligned} \tag{13}$$

Equation (13) is a general expression for the received carrier signal. An expression for the received response at the detector output, $g(t)$, can be obtained by considering only the envelope of this carrier. This allows $g(t)$ to be evaluated for any arbitrary attenuation or phase characteristic. In applying these results in the work to follow, the complete generality

of (13) will not be preserved, since a special type of transmission characteristic is being considered. By sacrificing some of the generality, however, an advantage is gained in the facility with which the results can be logically organized and utilized.

In evaluating the way a data system's performance varies as the transmission-frequency characteristics are altered, the results will be more lucid and systematic if the phase and attenuation are considered separately. While these two qualities are not independent in physical transmission facilities, it is not entirely academic to treat them independently, since this approach is often used in equalizing facilities.

In this study the influence of sinusoidal phase variation as given by (4) will be considered in some detail, and, in order to isolate the effects of this influence, the attenuation will be considered equal to unity over the frequency band of interest. (A similar study could be made of the effects of attenuation by specifying linear phase characteristics and allowing the attenuation to vary systematically.) For computational purposes we will use an approximate expression for $r(t)$ by considering a finite number of frequency components. Applying these conditions to (13) and rewriting this equation for notational convenience, the carrier response may be expressed as

$$\begin{aligned} r(t) &= (M + N) \cos \omega_0 t + (R + S) \sin \omega_0 t \\ &= [(M + N)^2 + (R + S)^2]^{\frac{1}{2}} \cos [\omega_0 t + \varphi(t)], \end{aligned} \tag{14}$$

where

in-phase components:

$$M = \sum_{n=0}^{N} C_n \cos \left[\frac{n2\pi}{T} t + b \sin \left(\frac{n2\pi}{T} \tau + \theta \right) \right],$$

$$N = \sum_{n=1}^{N} C_n \cos \left[\frac{n2\pi}{T} t + b \sin \left(\frac{n2\pi}{T} \tau - \theta \right) \right];$$

quadrature components:

$$\begin{aligned} R &= - \sum_{n=0}^{N} C_n \sin \left[\frac{n2\pi}{T} t + b \sin \left(\frac{n2\pi}{T} \tau + \theta \right) \right], \\ S &= \sum_{n=1}^{N} C_n \sin \left[\frac{n2\pi}{T} t + b \sin \left(\frac{n2\pi}{T} \tau - \theta \right) \right] \end{aligned} \tag{15}$$

and

$$\varphi(t) = \tan^{-1} \left(\frac{R + S}{M + N} \right). \tag{16}$$

In (14), the expression for the carrier response, the envelope of the carrier signal that is the output of the envelope detector is given by

$$g(t) = \sqrt{(M + N)^2 + (R + S)^2}. \tag{17}$$

The desired expressions, which relate the output of the detector $g(t)$ to the input signal $f(t)$ (as manifest in C_n), and the three parameters b, τ and θ, which describe the phase characteristics of the transmission path, are given by (15) and (17). It is these equations that form the basis of the computational work on the digital computer. A careful examination of them will give not only a valuable insight into the influence of sinusoidal phase characteristics on AM transmission, but will also serve as a guide in computational work and help in interpreting results. It is evident from (14) that the terms M and N may be considered as modulating a cosine carrier, and are, therefore, referred to as the *in-phase components* of the received signal, while the terms R and S modulate a sine carrier and are designated as *quadrature components.* The expression for the envelope response, (17), indicates that these two quantities cannot be superimposed directly, but must be added at right angles. Also it can be seen from (15) that for $b = 0$ the quadrature components cancel and the in-phase components add together, and the received signal is an exact replica of the transmitted signal. For $b \neq 0$ but $\theta = 0$, the quadrature terms cancel and the in-phase terms add to give a distorted replica of the transmitted signal at the receiver. For the case $b \neq 0$ and $\theta \neq 0$, the received signal is a distorted replica of the transmitted signal that contains both in-phase and quadrature components.

It can be seen from (3) and Fig. 3 that a reversal of the sign in b amounts to inverting the phase curve with respect to the linear phase line. It has been shown[1] that such a reversal in the phase curve results in a reversal in time of the received signal. This important fact will be utilized later in interpreting the performance results. Another useful observation that can be made from examining (15) is that a reversal in the sign of θ in these equations leaves the received signal unaltered.

V. COMPUTED PULSE RESPONSE

Equations (1), (15) and (17) provide a means for computing the output of the envelope detector of an AM transmission system subject to a sinusoidal variation in the phase characteristics, when the periodically recurring raised cosine pulse of Fig. 3 is applied at the input. These equations were programmed for numerical evaluation on the IBM 704 computer. In evaluating these equations, the computations were carried

out for a specific bit speed and specific values of b, τ and θ. The results from these computations, however, can be given on a normalized basis. To do this it is necessary to express the frequency scale of the phase or delay curves in terms of the bit speed rather than in cycles per second, and also to express the time scale of the response function and the delay curve in terms of the bit length rather than in seconds.

A number of normalized pulse responses are shown in Fig. 5, together with the normalized delay that produces the responses. The curves of Fig. 5(b) are for the case of symmetrical delay ($\theta = 0$) as shown in Fig. 5(a). For these cases, the quadrature component is zero, as would be expected from (15). The response curves in this figure are plots of the in-phase component of the received signal.

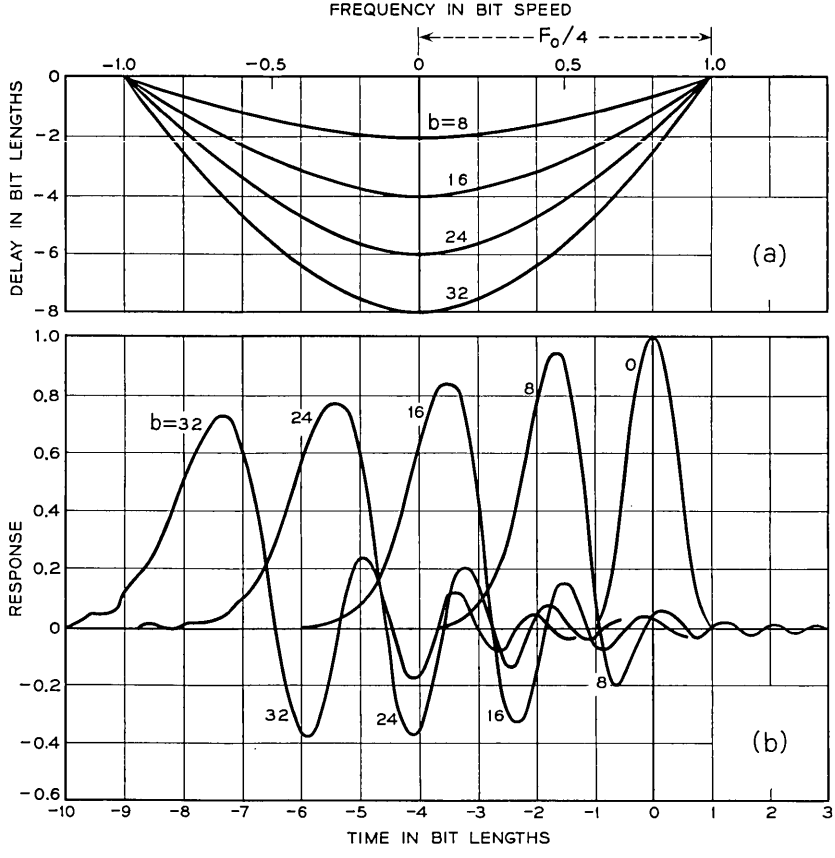


Fig. 5 – (a) Normalized delay; (b) corresponding pulse response.

To illustrate how these normalized results can be related to specific cases, consider an input pulse transmitted at a bit speed of 1200 bits per second (bit length = 833.3 microseconds). From Fig. 5(a) the period of the delay curve F_0 is equal to four times the bit speed or 4800 cycles, giving a value of $\tau = 1/F_0 = 208.33$ microseconds. The curve for $b = 8$ in Fig. 5(a) has a maximum delay excursion of $b\tau = 8 \times 208.3$ microseconds = 1666.3 microseconds = 2 bit lengths. The pulse response corresponding to this delay is shown in Fig. 5(b) for $b = 8$. These same results may also be applied to other bit speeds. For a bit speed of 2400 bits per second (bit length = 416.7 microseconds), $F_0 = 9600$, giving $\tau = 104.2$. Again the curve for $b = 8$ has a maximum delay of two bit lengths = $b\tau = 833.3$ microseconds, and the same pulse response is produced here as for the low bit speed case.

In Fig. 6 pulse response curves are shown for conditions of asymmetry in the delay characteristics. This gives rise to the quadrature component in the received signal, and the resulting effect on the envelope response is shown. Here again the curves are normalized with respect to bit length. For the curve of Fig. 6(a), $\theta = \pm 45°$; for Fig. 6(b), $\theta = \pm 90°$. These curves illustrate the fact pointed out previously: that the response is independent of the sign of the angle θ.

These distorted pulse response curves indicate a degradation in transmission manifest in a general lowering of the pulse peak value and a spreading of the pulse in time. For the cases of larger distortion this results in a considerable amount of intersymbol interference between adjacent pulses in a data signal. This, however, is only a qualitative picture of the influence of delay distortion on system performance. In order to make this picture more precise, it is necessary to consider a signal consisting of a number of pulses making up a pattern such as that illustrated in Fig. 2. This extension to the basic operation of computing pulse response was achieved by combining the received response of a single pulse (before envelope detection) in accordance with the desired pattern. In doing this, the in-phase components from adjacent pulses were added together. Similarly, the quadrature components were added together, and the resulting envelope response was determined in accordance with

$$g(t) = \sqrt{[\Sigma(M + N)]^2 + [\Sigma(R + S)]^2}, \tag{18}$$

where M, N, R and S are defined by (15). By this method, the results previously discussed can now be extended so that the output response for pulse patterns can be obtained.

An example of pulse-pattern response showing the resultant in-phase

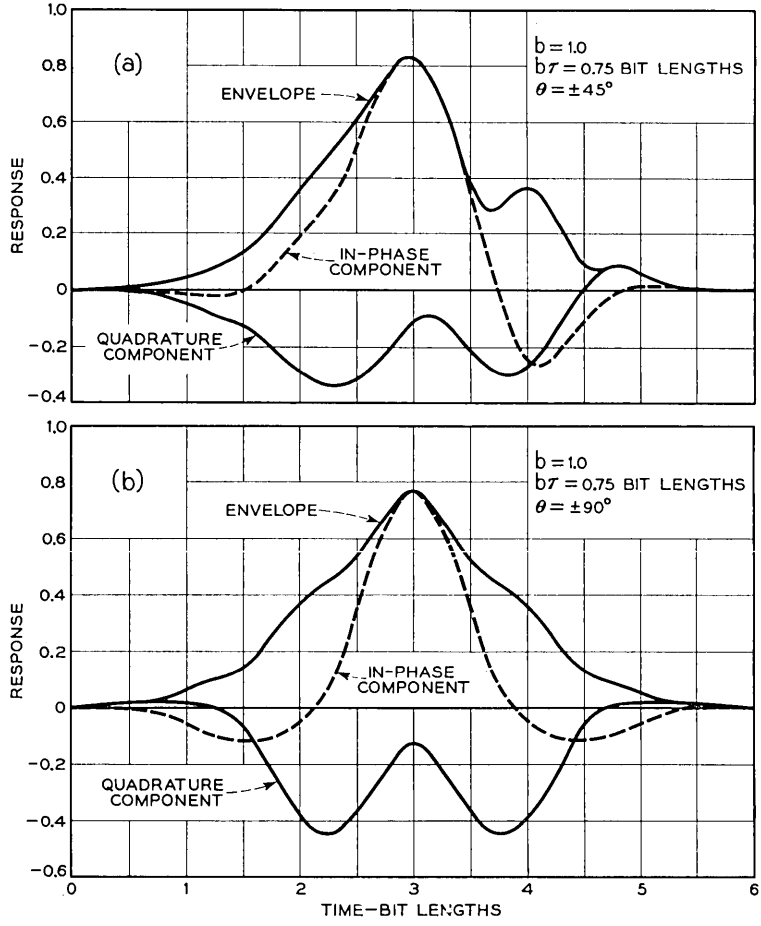


Fig. 6 – Pulse response curves.

and quadrature components and the envelope response is given in Fig. 7. In this figure a pattern of binary characters is shown together with the corresponding output response curve for conditions $b = 1$, $b\tau = 0.75$ bit length and $\theta = 90°$. In order to recover the binary information from the signal, the envelope response is examined at the sampling-time points to determine if the value of the curve at each of these points is less than or greater than the value of the slice level. The distance from the signal to the slice level for a given time point (as indicated by the arrows) is a measure of the noise margin for that point. The amount by which the noise margin is reduced by the transmission characteristic can be de-

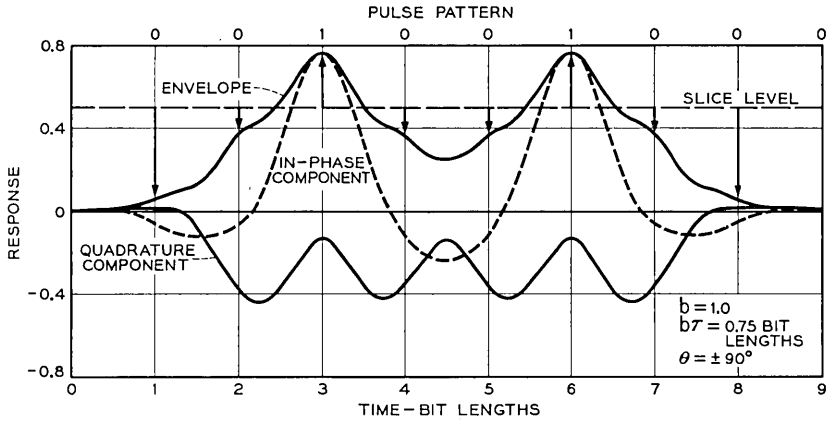


Fig. 7 – Basic method for evaluating performance.

termined by comparing these values with the corresponding values for an undistorted or reference output resulting from ideal transmission characteristics.

This procedure makes possible a quantitative evaluation of data system performance. It is evident, however, that for longer pulse patterns this procedure would soon become tedious and impractical if carried out in the manner shown. In order to alleviate this difficulty, the basic method for evaluating performance, as illustrated in Fig. 7, was included in the computer operation for any desired pulse pattern up to a certain maximum length. As the delay distortion increases, the range of intersymbol interference of a single pulse becomes larger; it was found necessary in this analysis to include cases for which the intersymbol interference extended over as many as nine bit lengths. In order to take this influence into account it was necessary to use a pattern that contained all possible combinations of binary characters that could occur in a nine-bit word. This results in a maximum pattern length of $2^9 = 512$ bits. A binary pattern of this type was used in the computer simulation work that was carried out.

VI. EYE PATTERN

One way of portraying this procedure for evaluating data system performance for a long sequence of pulses is by means of the "eye" pattern. The manner in which this pattern is made can be illustrated by dividing the output response signal from the system into a number of segments each of duration equal to two bit lengths and plotting all these

segments over the same two-bit time intervals. Such a plot for a reference or undistorted eye is shown in Fig. 8(a). Here the eye pattern is made up of traces that are straight lines at the top and bottom of the eye, corresponding to a sequence of ones and zeros, respectively, and traces that have sinusoidal shape in between, representing transitions from one to zero and zero to one. Since the pulses are undistorted, all traces of a particular type fall on top of one another, and the resulting eye pattern is made up of well-defined lines. In this case, the opening or aperture A, of the eye is equal to A_0, which is the maximum value of a single pulse. The quantity A is a measure of the noise margin and will be used as a criterion for data system performance.

For conditions of transmission distortion, the resulting eye pattern is illustrated by Fig. 8(b). Here the various individual traces have been distorted and no longer form well-defined lines of transition, but are spread out into bands. This results in a reduction of the aperture, which can be expressed numerically as an impairment in the system performance:

$$\text{transmission impairment} = 20 \log_{10} \frac{A_0}{A}. \tag{19}$$

There are two factors that determine the validity with which such an

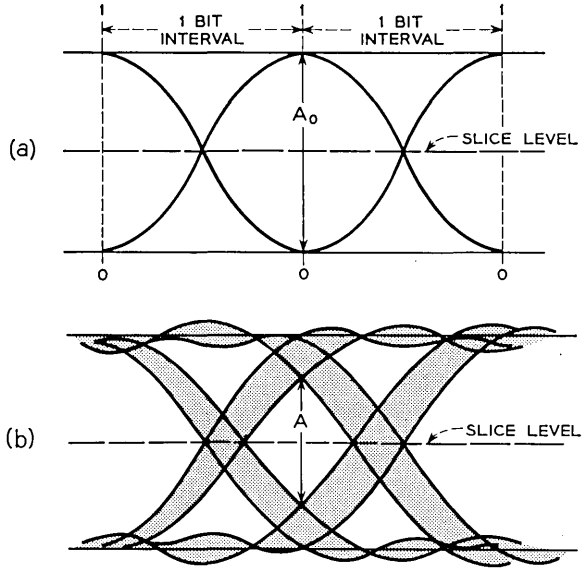


Fig. 8 – (a) Reference eye pattern; (b) received eye pattern with transmission distortion.

expression of performance describes an actual system. First, it must be acknowledged that the effective opening of the eye may change from one trace to another, and the quantity A as shown in Fig. 8(b) is a minimum bound of this opening. Since an opening of this magnitude will occur relatively infrequently this quantity is then a pessimistic estimate of the system performance; getting an exact measure would involve taking account of the statistical distribution of the quantity A. The probability of an error occurring would then be determined by convoluting this distribution with the distribution of noise. Such a refinement is rather complicated, and may not be necessary except in cases of very small aperture, since, for the larger values of A, the traces across the eye are rather densely packed.

The other fact to be considered is that in this simulation the degradation in performance due to timing recovery is not included. These results then specify the performance under conditions of ideal timing recovery, or for systems operating with ideal transmitted timing. Again, this is a good approximation to system operation with timing recovery, except for small aperture values, where timing jitter is unavoidable.

VII. EXPERIMENTAL VERIFICATION OF COMPUTER SIMULATION

It is of interest to determine how closely the performance results from this simulation procedure check with measured performance from an actual data system. The aperture value of an AM double-sideband data system operating in the voice band frequency range was measured at several bit speeds over a given transmission channel. The transmission characteristics of the channel were then applied to the computer simulation of the data system for the same bit speeds. The results from these two tests expressed in terms of aperture of the eye pattern are shown in Fig. 9. The fact that the computer simulation performance is slightly better over the range of bit speeds tested is to be expected, because of the assumption of idealization in the modulator and detector in making the simulation. These results tend to confirm the validity of the simulation technique as a means for evaluating data system performance.

VIII. RESULTS

A criterion for data system performance has been specified as A, the aperture of the eye pattern formed by the received signal of a data system. The results of computing performance based on the data system simulation are presented here as a portrayal of the manner in which A varies with a systematic variation of the parameters describing the

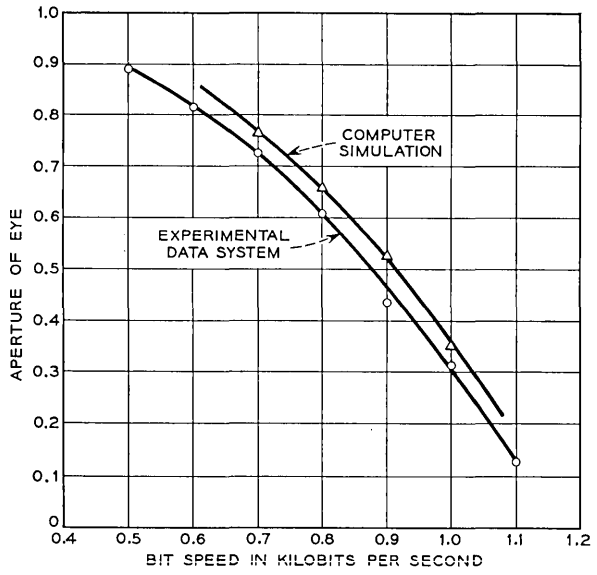


Fig. 9 – Comparison of computer simulation performance with experimental verification.

transmission characteristics of the system. To make these results more general, a normalized parameter δ, will be defined as $\delta = \tau \times$ bit speed. Results showing the manner in which the quantity A varies with b, δ and θ are given in Figs. 10 through 20, where b is expressed in radians and θ in degrees and δ is a quantity equal to the reciprocal of the period of the sinusoidal phase characteristics normalized with respect to bit speed.

In these figures, performance curves are plotted with the dependent variable b ranging over values sufficient to cause the aperture to vary from a reference value of unity to complete closure. The quantity δ varies from 0.2 bit length to 1.4 bit lengths. The degree of asymmetry is specified by the angle θ, which varies as a parameter from zero to 90 degrees. It is an apparent feature of each set of curves that the aperture value is reduced more rapidly, as b increases, for asymmetric phase conditions than for symmetrical conditions. This is an evidence of the influence of the quadrature component on data system performance. From these curves, it can also be seen that the relative influence of the quadrature component is much more pronounced for small values of δ than for larger values.

While the performance curves cover only one quadrant of the sinusoi-

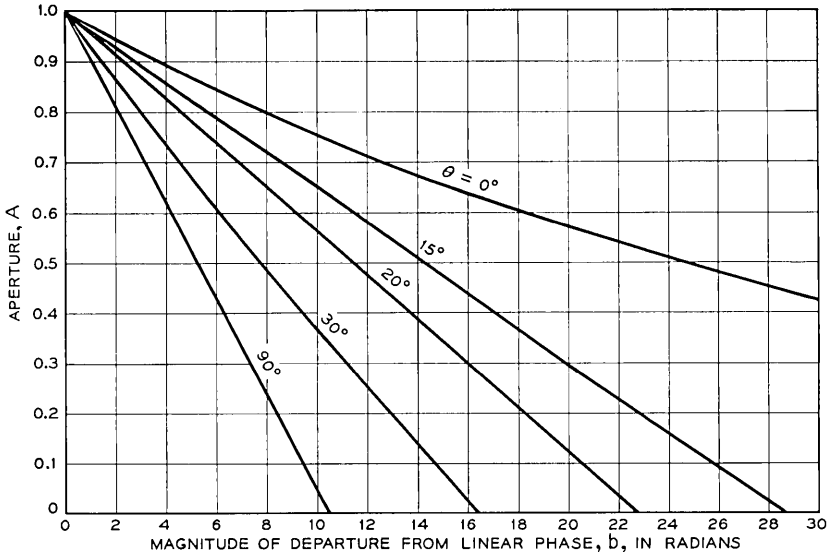


Fig. 10 – Plot of aperture vs. b for $\delta = 0.2 = \tau \times$ bit speed.

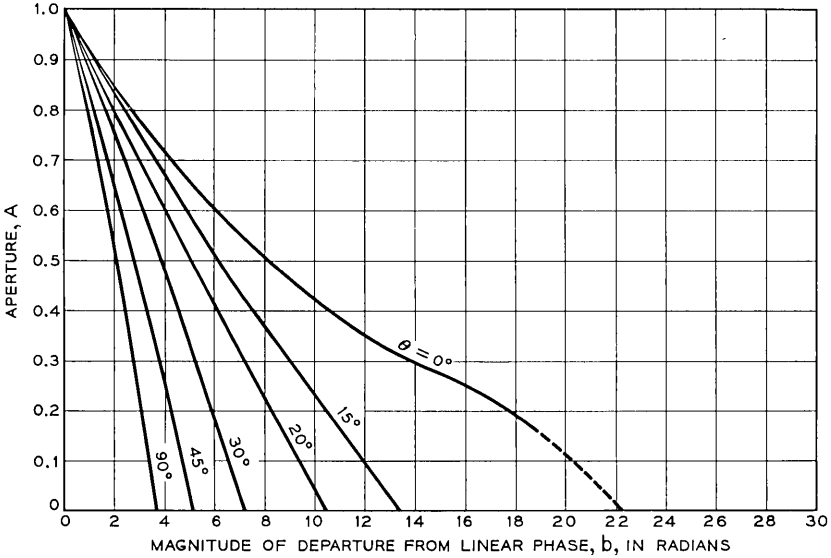


Fig. 11 – Plot of aperture vs. b for $\delta = 0.3 = \tau \times$ bit speed.

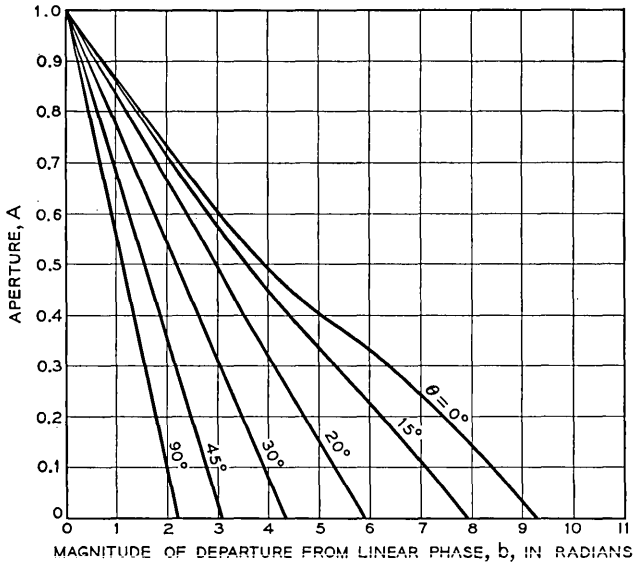


Fig. 12 – Plot of aperture vs. b for $\delta = 0.4 = \tau \times$ bit speed.

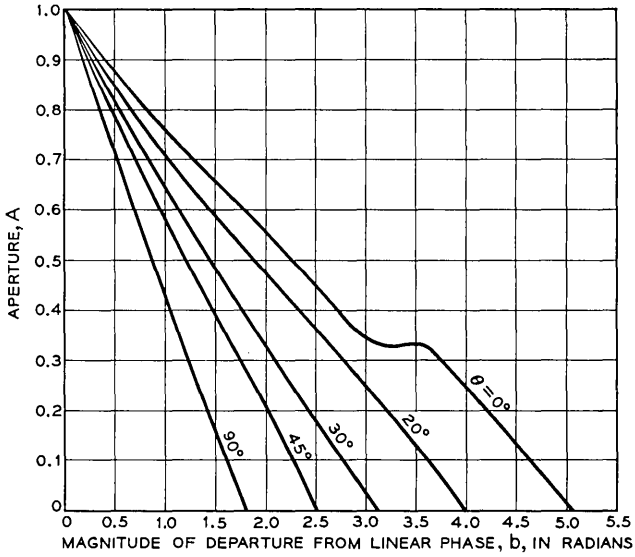


Fig. 13 – Plot of aperture vs. b for $\delta = 0.5 = \tau \times$ bit speed.

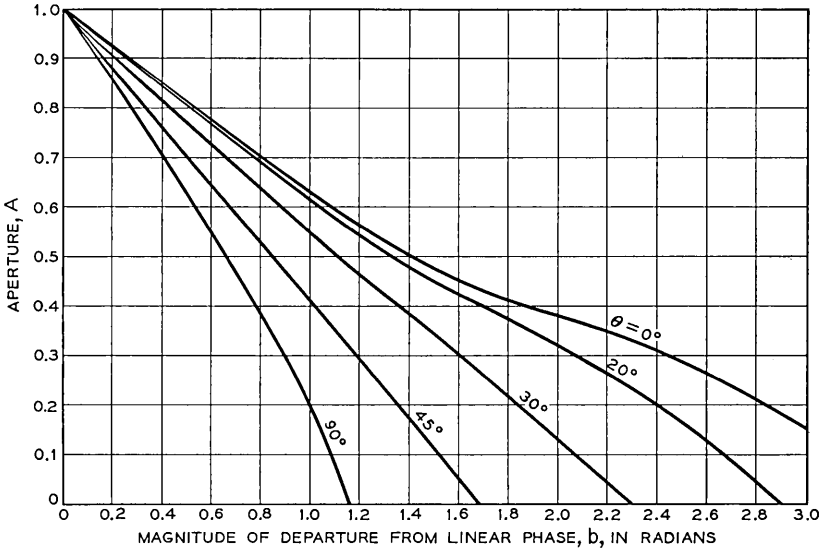


Fig. 14 – Plot of aperture vs. b for $\delta = 0.6 = \tau \times$ bit speed.

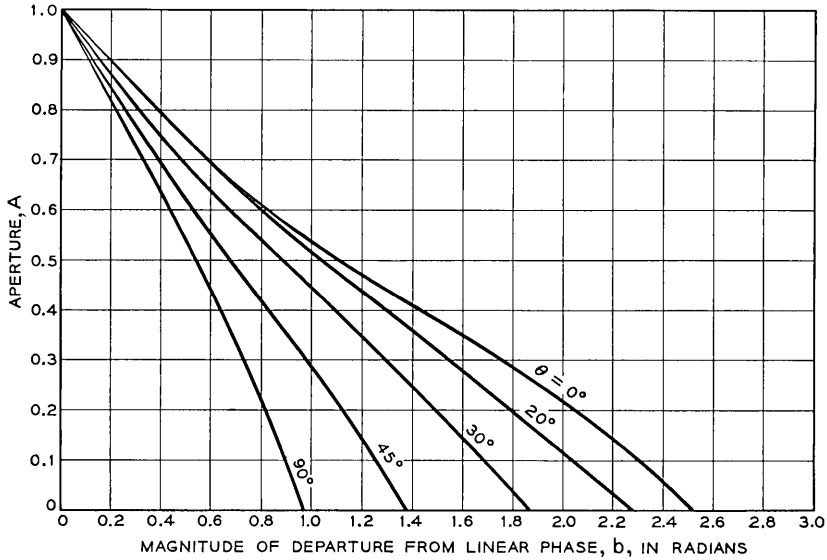


Fig. 15 – Plot of aperture vs. b for $\delta = 0.7 = \tau \times$ bit speed.

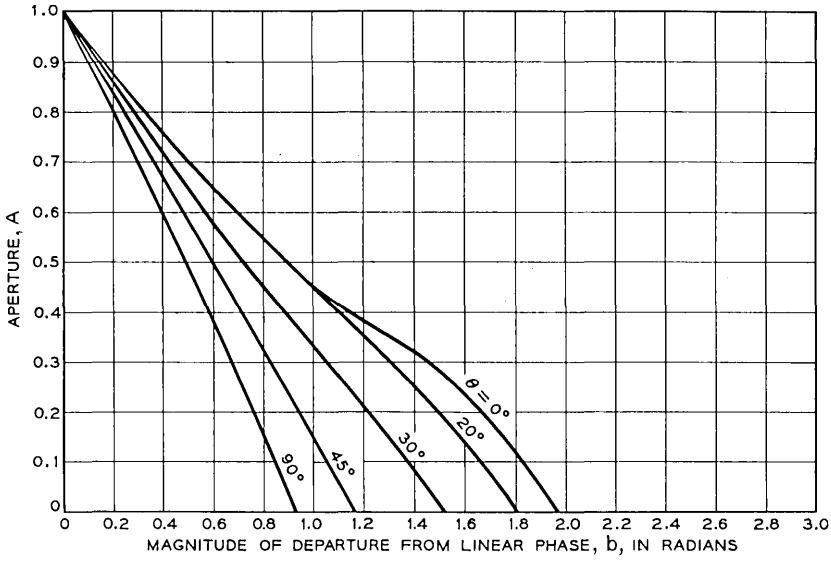


Fig. 16 – Plot of aperture vs. b for $\delta = 0.8 = \tau \times$ bit speed.

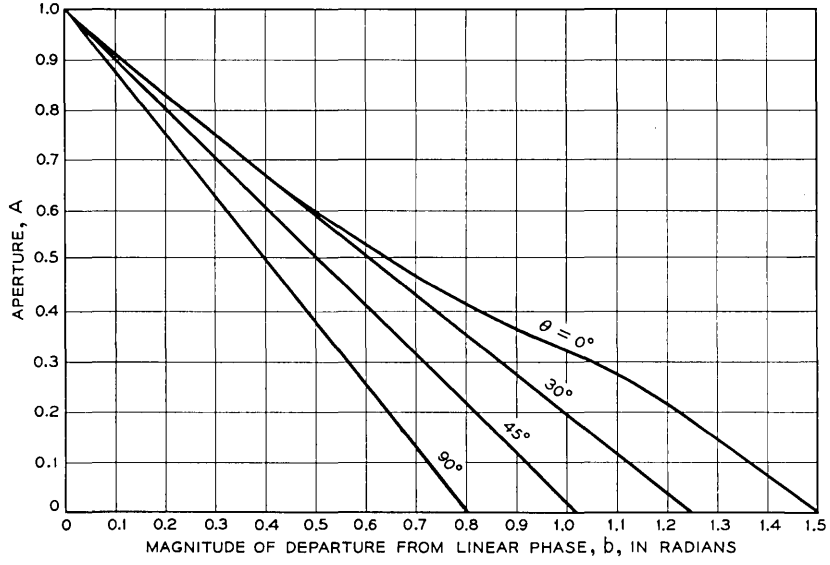


Fig. 17 – Plot of aperture vs. b for $\delta = 0.9 = \tau \times$ bit speed.

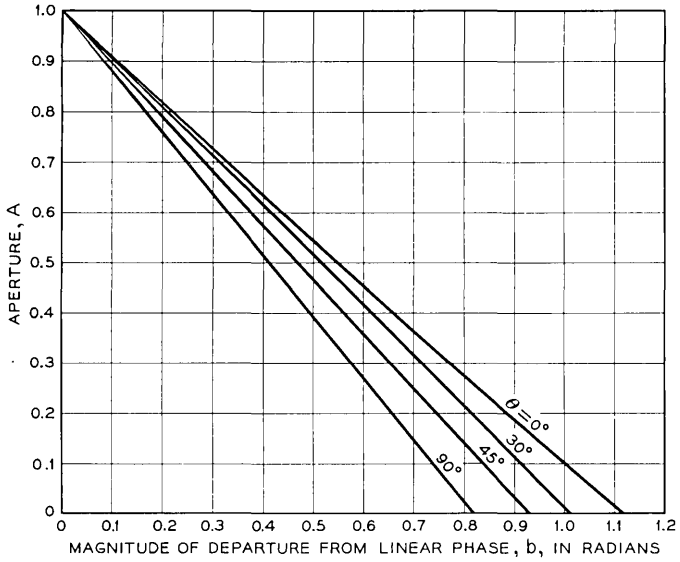


Fig. 18 – Plot of aperture vs. b for $\delta = 1.0 = \tau \times$ bit speed.

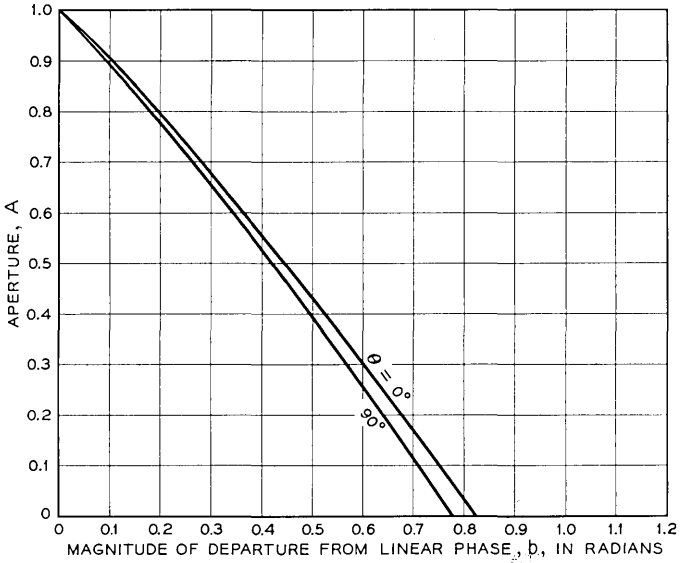


Fig. 19 – Plot of aperture vs. b for $\delta = 1.2 = \tau \times$ bit speed.

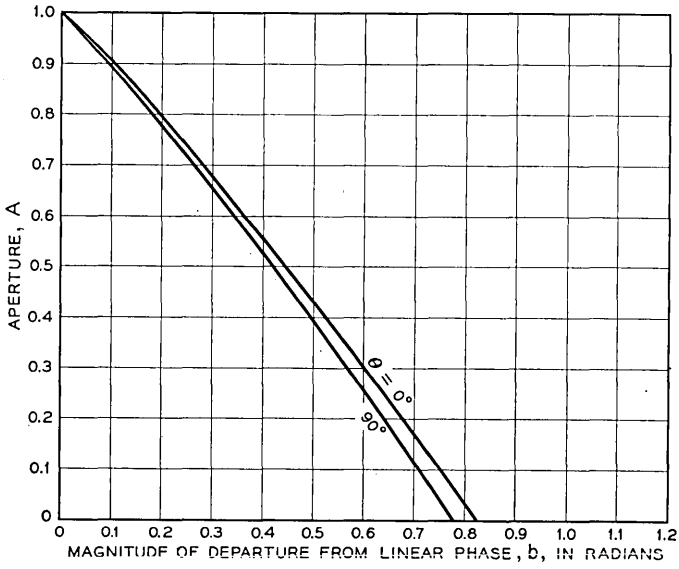


Fig. 20 – Plot of aperture vs. b for $\delta = 1.4 = \tau \times$ bit speed.

dal phase curve, the other three quadrants are duplicates of the curves shown. This can be seen from the facts previously stated, which indicated that a change in the sign of θ leaves the output response unchanged, and that a change in the sign of b reverses the time variable of the response and therefore reverses time in the eye pattern but does not change the aperture value.

The information given on the performance curves of Figs. 10 through 20 provides a means for evaluating the aperture of a data system for a wide range of sinusoidal phase characteristics, either by direct means or by interpolation. These results can be displayed in another manner, which gives a more concise picture of how the performance varies. To do this the performance information for a particular value of δ is plotted in polar form with b and θ as the radial and angular coordinates respectively. Such a plot is shown in Fig. 21 for $\delta = 0.75$, with contours of constant aperture shown as a parameter. The outer contour has a value of $A = 0$ and is the locus of points for which the aperture is completely closed. Any sinusoidal phase characteristic whose period corresponds to this value of δ can be located on the polar diagram of Fig. 21 as a point. If the magnitudes of b and θ are such that the point lies within the outer contour, the aperture value can be determined either directly or by interpolation. It is now possible to plot a path on this diagram showing how the value of A varies for changes in the values of b and θ.

This type of plot can be made for other values of δ. By combining such plots, the information given in Figs. 10 through 20 results in a very useful three-dimensional model with coordinates b, δ and θ, which depicts the performance of the data system. Such a model is shown diagrammatically in Fig. 22. In this three-dimensional space any sinusoidal phase (or delay) curve is represented as a point. A number of curves that illustrate the particular delay that occurs at various points are shown by Fig. 22(a), (b), (c) and (d). The surface of the model indicated by the dashed line in Fig. 22 represents the locus of points for which the aperture

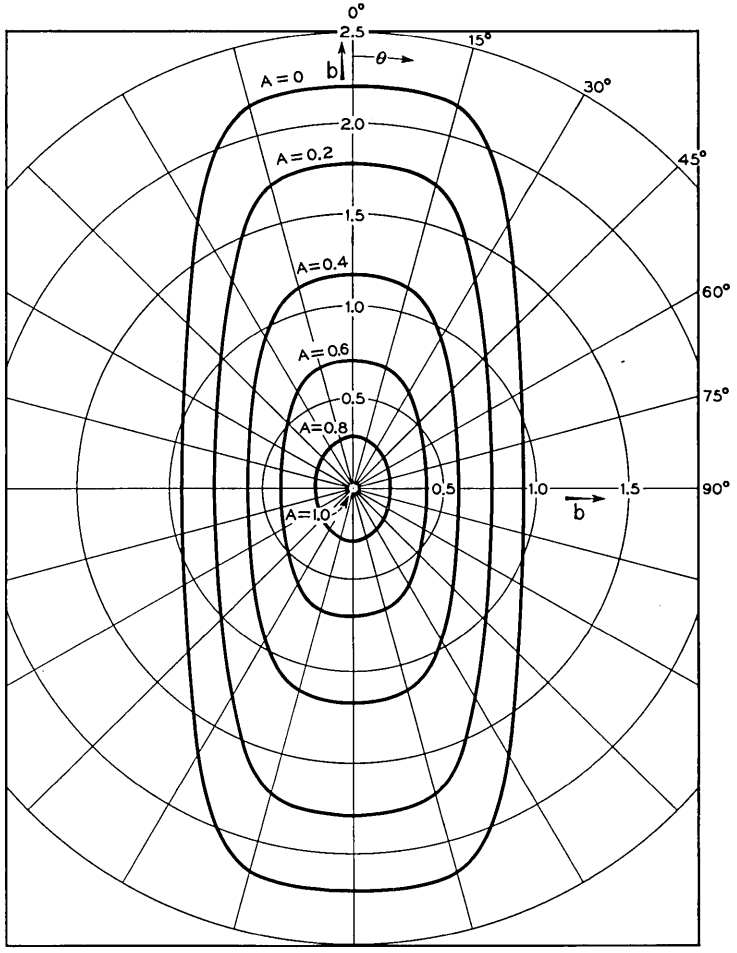


Fig. 21 – Polar diagram of b vs. θ for $\delta = 0.75$ with aperture as a parameter.

value is zero. For points inside this surface the aperture is greater than zero. Other surfaces within this model can be described by connecting all points having a common value of A. The points lying along the vertical axis represent conditions of zero distortion, or maximum aperture value A. A photograph of a performance model constructed from the numerical results obtained from the computer simulation is shown in Fig. 23.

IX. APPLICATION OF RESULTS TO ACTUAL TRANSMISSION FACILITIES

In the simulation of the data system that has been presented the phase characteristics have been assumed to be sinusoidal in shape, for reasons of mathematical convenience. Some rather general results have been given in terms of the parameters that describe the phase characteristic. In this section several specific transmission facilities will be considered in the light of this simulation technique.

The validity with which the sinusoidal phase (or cosine-shaped delay) represents actual transmission characteristics will be investigated, and computed performance resulting from data system operation over these facilities is given.

In Fig. 24, the delay characteristics of four message facilities are shown by the solid lines, together with a cosine approximation to these delay curves given by the dashed lines. It can be seen from Fig. 24 that the cosine-shaped delay curves give a very good approximation to the actual

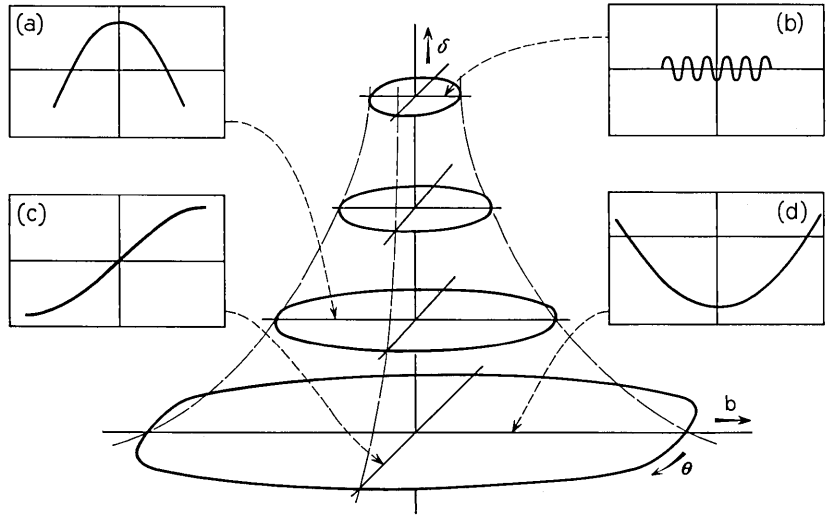


Fig. 22 – Three-dimensional model of data system performance.

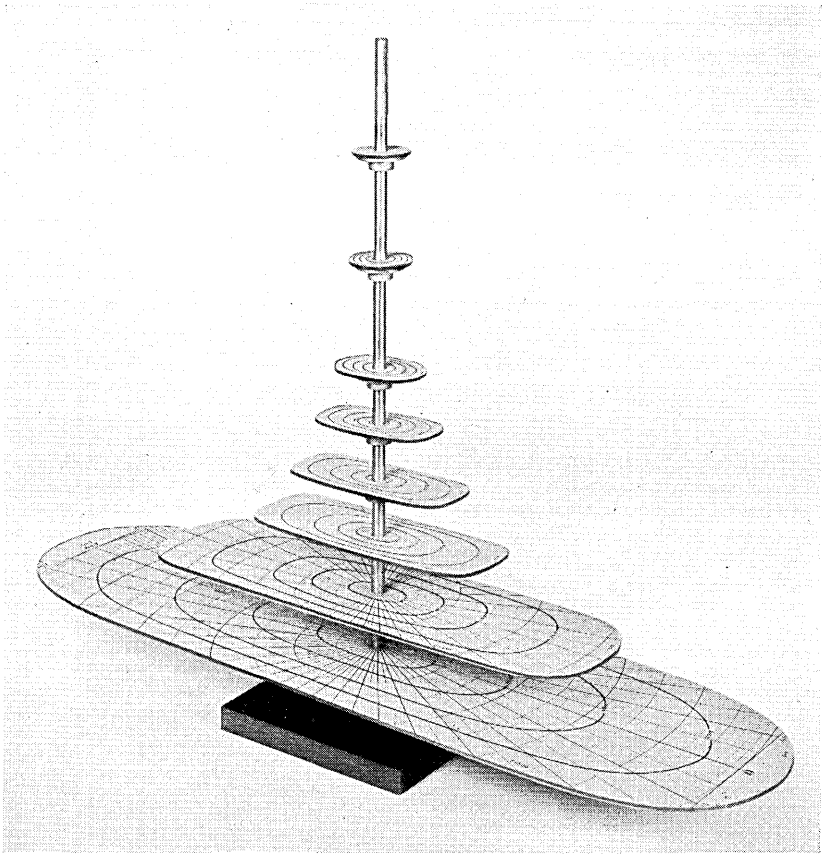


Fig. 23 – Performance model constructed from numerical results of computer simulation.

delay over the frequency range of 1000 to 2600 cycles per second, the maximum departure of the approximating curve from the actual facilities being less than 100 microseconds over this frequency range. The approximating curves can then be specified by the three parameters b, τ and θ (or when the bit speed of the data system is given by b, δ and θ). These values are given in Table I, for the cases of a data system operating with an 1800-cycle carrier, and at bit speeds of 1200 and 1500 bits per second. The resulting aperture values and transmission impairment figures produced are also given in Table I. Aperture values were determined by direct computer computation; these values can also be determined to a good approximation by interpolation from Figs. 10 and 11.

The method shown here for determining the impairment values in

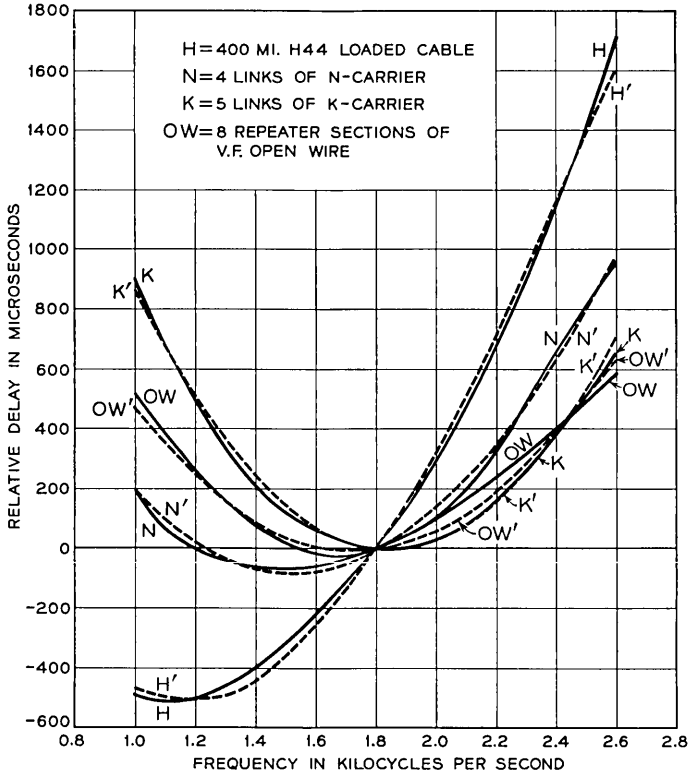


Fig. 24 – Delay characteristics of four message facilities.

Table I illustrates how data system performance can be established for actual transmission facilities from the results of the computer simulation. This procedure can also be used to determine how performance varies with some of the parameters that are important in the design and engineering of data systems, such as bit speed, location of carrier frequency and length and type of transmission facility. For example, as the bit speed is increased from 1200 to 1500 bits per second, the impairments of the facilities are increased by the amount shown. (The loaded cable changes from marginal operation to complete failure.)

The poor performance of the loaded cable facility is due to the fact that the location of the carrier frequency results in an asymmetrical disposition of the delay curve about the carrier, resulting in a considerable amount of quadrature component. If the location of the carrier were changed from 1800 to 1500 cycles per second, the angle θ would be

TABLE I

Facility	b, radians	τ, μ sec	θ, degrees	1200 bits per second			1500 bits per second		
				δ, bit lengths	A^*	Transmission impairment, db	δ, bit lengths	A^*	Transmission impairment, db
400-mile H44 loaded cable	9.26	194	43.8	0.233	0.092	20.7	0.292	—	—
4 links of N carrier	11.8	163	15.8	0.196	0.60	4.4	0.244	0.38	8.4
5 links of K carrier	9.89	190	−3.1	0.228	0.67	3.5	0.285	0.50	6.0
8-repeater sections of V.F. open wire	11.2	161	3.6	0.193	0.76	2.4	0.242	0.61	4.7

* These aperture values were determined by direct computation. They can also be obtained by interpolation from Fig. 10 and Fig. 11, with the exception of the loaded cable (which has an aperture value of either zero or very nearly zero).

changed from 43.8° to 22.8°. For a bit speed of 1200 bits per second, this gives an aperture value of 0.41 (by interpolation from Figs. 10 and 11) and a resulting transmission impairment of 7.8 db.

Performance for many other transmission conditions of interest can be determined with similar ease by this same procedure.

X. CONCLUSIONS

It is apparent that the data system simulation technique that has been developed and used here for evaluating performance is a powerful and valuable tool. Wherever possible, experimental and theoretical verifications of the results obtained from this simulation have been made. These tests indicate that this method is valid and accurate. The performance curves given have been used to illustrate the ease and effectiveness with which many of the complex problems relating to delay distortion can be answered in a rather general way.

Several uses can be made of this simulation technique. First, it provides a means of obtaining quantitative answers to a broad category of problems relating to data transmission, such as the influence on performance of the variation of bit speed, the location of carrier frequency and the length and type of transmission facility used. In addition, a performance model based on a sinusoidal phase characteristic and normalized with respect to bit speed has been produced for the data system under consideration. The surface of this model describes the boundary (in terms of variation of phase distortion) within which the system can operate. Similar models can be established for other types of data systems to provide a means for comparing the performance on a much broader basis than is presently possible.

The investigation presented here was directed to the problem of determining what influence phase (or delay) distortion has on data system performance. It is clear that the simulation method developed could be extended to include other problems of data transmission, such as a general study of the influence of the attenuation characteristics, the problem of recovering timing and the influence of noise on system transmission. Additional work in this general area is now being carried out.

XI. ACKNOWLEDGMENTS

The author wishes to acknowledge the assistance received in carrying out this work. The simulation program for this study was worked out in collaboration with Miss Geraldine Nelson, who carried out the many programming details. Much of the actual computational work on the computer and many other details involved in the preparation of this material were carried out by Miss Juanita Sacks.

REFERENCES

1. Fowler, A. D., and Gibby, R. A., Assessment of Effects of Delay Distortion in Data Systems, Comm. & Electronics, no. 40, 1959, p. 918.
2. Sunde, E. D., Theoretical Fundamentals of Pulse Transmission, B.S.T.J., **33**, 1954, p. 721; 987.
3. Sunde, E. D., Ideal Binary Pulse Transmission by AM and FM, B.S.T.J., **38**, 1959, p. 1357.
4. Oliver, B. M., Pierce, J. R. and Shannon, C. E., The Philosophy of PCM, Proc. I.R.E., **36**, 1948, p. 1324.
5. Mertz, P., Transmission Line Characteristics and Effects on Pulse Transmission, in Proc. Symp. on Information Networks, Polytechnic Inst. of Brooklyn, Brooklyn, N. Y., 1954.

Semiconductor Strain Transducers

By F. T. GEYLING and J. J. FORST

(Manuscript received February 9, 1960)

The relatively recent discovery of the high piezoresistive sensitivity of semiconductors such as germanium and silicon has made miniature strain transducers available that compare with piezoelectric devices in dynamic applications but are also capable of producing dc signals After a historical survey of research in piezoresistivity, a brief phenomenological description of the physical effect is given and the pertinent solid state theory is summarized. Various applications of semiconductor elements as strain transducers are then discussed in detail, special emphasis being given to surface strain gages. The problems of maximizing the bond rigidity for such gages and compensating for temperature effects are dealt with, and future efforts in these directions are outlined.

I. INTRODUCTION

Around the end of 1958, considerable interest developed among experimenters and transducer manufacturers in the possible application of highly sensitive piezoresistive elements as subminiature sensing devices for displacements, strains and forces. For dynamic problems, piezoelectric pickups consisting of quartz or barium titanate had previously found their way into numerous applications for the measurement of vibrations, shock intensities and stress waves. However, similar lightweight instruments that could handle dc or near-dc signals as well as dynamic ones were yet to be found. Piezoresistive semiconductor materials do just that. The present paper is intended to give a summary of established facts about them and indicate some of their potentialities.

The history of research in piezoresistive phenomena dates back to the 1920's, when some of the earliest observations were made by Bridgman on single and polycrystalline metal specimens.[1,2] These results were augmented by several contributions in the 1930's from Allen and Cookson,[3-7] who measured the effects that were subsequently utilized in wire strain gages. To the authors' knowledge, the first piezoresistive measure-

ments on semiconductors were made in 1953 in connection with basic research for the transistor technology.[8,9] From this point on, further investigations of piezoresistance in semiconductors were carried on very actively in an experimental direction[10-15] and along theoretical lines.[13,16,17] Shortly after the original measurements, efforts were made toward applications of the phenomenon, and these have continued ever since as one of the device technologies that evolved from the original work in transistor development.[18-26]

We begin this review with a simplified account of the essential solid state theory and its experimental corroborations. This sketch of existing knowledge is intended for the experimentalist concerned with strain transducers and does not make any claims regarding novelty. It merely outlines the basic thoughts and gives references for additional detail. The remainder of the paper covers several applications of piezoresistive transducers at room temperature, and then proceeds to a discussion of strain gage applications at various temperatures.

The authors are indebted to W. P. Mason, C. Herring, R. N. Thurston, R. O'Regan, J. S. Courtney-Pratt, and W. L. Feldmann for help and suggestions, and to several previous authors in this field for permission to use some of the material listed in the references at the end of this paper.

II. PIEZORESISTIVE PROPERTIES OF SEMICONDUCTORS

2.1 *Phenomenological Description*

From routine applications of Ohm's law, we are used to thinking of the relation between potential difference and current as involving only a scalar constant of proportionality, the resistance, between the vectors **V** and **I**. This is trivially obvious for a unidimensional conductor such as a link in some circuit. If we restate the situation for conduction in a three-dimensional medium, however, then the current density vector, **i**, generated by a potential gradient, **E**, will not in general have the same direction as the latter unless the medium is isotropic or cubic; i.e., $\mathbf{i} = (1/\rho)\mathbf{E}$ if the resistivity ρ is the same in all directions within the conductor. For monocrystalline conductors this is not necessarily the case; i.e., the current vector resulting from an impressed voltage gradient is not colinear with the latter and, conversely, if some specified **i** is maintained in the material the attendant **E** is not colinear with it. We may formulate this situation more concisely by identifying the components of **E** and **i** with numeric subscripts indicating the directions of orthogonal crystal axes. Then $E_1 = \rho_{11}i_1 + \rho_{12}i_2 + \rho_{13}i_3$, etc., where the first sub-

script for each ρ_{ij} indicates the field component it is contributing to, and the second subscript identifies the current component making the contribution. In literal notation and using the summation convention, we have

$$E_i = \rho_{ij} i_j , \tag{1}$$

where all subscripts have the range 1 to 3. Only if $\rho_{ij} = 0$ for $i \neq j$ and $\rho_{11} = \rho_{22} = \rho_{33} = \rho$ does the medium provide isotropic conduction. Such is indeed the case for unstressed germanium and silicon, which possess a cubic crystal structure.

We now generalize the relation (1) to allow for piezoresistive effects due to a set of stresses T_{kl}, where T_{11}, T_{22}, T_{33} are normal stresses along crystal axes and T_{12}, T_{13}, T_{23} the shear stresses in this coordinate system. Then,

$$E_i = \rho_{ij} i_j + \pi_{ijkl} i_j T_{kl} , \tag{2}$$

where $\pi_{ijkl} T_{kl}$ are the stress-dependent contributions to the resistivity. The proportionality constants π are characterized not only by i and j, as was the zero-stress resistivity, but also by k and l, which relate each of them to a particular stress component T_{kl}. Clearly, (2) could be treated as a tensor relation, but, since we do not expect to carry out many operations on it during the following discussion, not much would be gained from this formalism. We shall, however, find it convenient to use the following contractions for various combinations of subscript values:

$$11 \sim 1,$$

$$22 \sim 2,$$

$$33 \sim 3,$$

$$23 \sim 4,$$

$$13 \sim 5,$$

$$12 \sim 6.$$

Utilizing the fact that the zero-stress resistivity for germanium and silicon is isotropic, we now introduce a new set of piezoresistance coefficients π_{st} such that, for example,

$$\rho\pi_{11} = \pi_{1111} ; \qquad \rho\pi_{12} = \pi_{1122} ; \qquad \rho\pi_{44} = 2\pi_{2323} .$$

If one finally observes all the symmetry conditions and the vanishing of certain π_{st},[20] one finds the following array for them:

$s\backslash t$	1	2	3	4	5	6
1	π_{11}	π_{12}	π_{12}	0	0	0
2	π_{12}	π_{11}	π_{12}	0	0	0
3	π_{12}	π_{12}	π_{11}	0	0	0
4	0	0	0	π_{44}	0	0
5	0	0	0	0	π_{44}	0
6	0	0	0	0	0	π_{44}.

(3)

The explicit form of (2) then becomes

$$\begin{aligned}
\frac{E_1}{\rho} &= i_1[1 + \pi_{11}T_1 + \pi_{12}(T_2 + T_3)] + \pi_{44}(i_2T_6 + i_3T_5),\\
\frac{E_2}{\rho} &= i_2[1 + \pi_{11}T_2 + \pi_{12}(T_1 + T_3)] + \pi_{44}(i_1T_6 + i_3T_4),\\
\frac{E_3}{\rho} &= i_3[1 + \pi_{11}T_3 + \pi_{12}(T_1 + T_2)] + \pi_{44}(i_1T_5 + i_2T_4).
\end{aligned} \tag{4}$$

We note that the second terms in the right-hand sides, $\pi_{11}T_1i_1$, $\pi_{11}T_2i_2$ or $\pi_{11}T_3i_3$, represent the piezoresistive effect as we know it from wire and foil gages. It is the effect of a stress in the direction of current flow on the potential drop in that direction. The additional terms in (4) simply reflect the more complicated behavior of the stressed lattice, which exhibits piezoresistive "carry-over" into E_i from stress components other than T_i.

In many applications of semiconductor transducers one is indeed concerned with the relation between a uniaxial state of stress T' and i' and E' in the same direction, where the latter, however, is not necessarily along a crystal axis, but exhibits the direction cosines l, m, n with respect to 1, 2, 3. The desired relation, based on the fundamental ones in (4), follows from

$$\begin{aligned}
i_1 &= li', \qquad & i_2 &= mi', \qquad & i_3 &= ni';\\
T_{11} &= l^2T', \qquad & T_{22} &= m^2T', \qquad & T_{33} &= n^2T';
\end{aligned}$$

and

$$T_{12} = lmT', \qquad T_{13} = lnT', \qquad T_{23} = mnT';$$

which we substitute into (4). Using the results in

$$E' = lE_1 + mE_2 + nE_3,$$

we can obtain

$$\frac{E'}{\rho} = i'\{1 + T'[\pi_{11} + 2(\pi_{44} + \pi_{12} - \pi_{11})(l^2m^2 + l^2n^2 + m^2n^2)]\} = i'[1 + \pi_l T']. \tag{5}$$

If $(\pi_{44} + \pi_{12} - \pi_{11}) \neq 0$ and π_{11} and $(\pi_{44} + \pi_{12} - \pi_{11})$ have the same sign, or if $|\pi_{44} + \pi_{12} - \pi_{11}| > 3|\pi_{11}|$, it can be shown that $|\pi_l|$ exhibits maxima at $l = m = n = \pm\sqrt{\frac{1}{3}}$, i.e., along the [111] axes. Otherwise, the maxima occur along the crystal axes.

Besides uniaxial stress in the [111] direction, we shall be interested in the effects of shear stresses T_4, T_5, T_6 alone. We observe from (4) that they produce a piezoresistive effect by virtue of π_{44}. In the event of hydrostatic pressure, finally, we note that $T_1 = T_2 = T_3 = -p$ and $T_4 = T_5 = T_6 = 0$, so that

$$\mathbf{E} = \rho\mathbf{i}[1 - p(\pi_{11} + 2\pi_{12})]. \tag{6}$$

Some typical values for the adiabatic piezoresistive coefficients of germanium and silicon at room temperature are given in Table I.[9]

The dimensions for the π follow from (4). $(m_l)_{111}$ is a so-called elastoresistive coefficient, which is computed from $(\pi_l)_{111}$ by means of Young's modulus in the [111] direction. Since it represents $\Delta\rho/\rho\epsilon$, where ϵ is the strain in the [111] direction, it is analogous to the "gage factor" defined for strain gages of the wire or foil type. We note that all these factors in the above table are considerably larger than a typical value of 2 for commercially available strain gages. This explains the current interest

TABLE I

Material	ρ, in ohm-cm	π, in 10^{-12} cm² per dyne				$(m_l)_{111}$, dimensionless
		π_{11}	π_{12}	π_{44}	$(\pi_l)_{111}$	
Germanium, n-type..	16.6	−5.2	−5.5	−138.7	−101.2	−157
Germanium, p-type..	1.1	−3.7	+3.2	+96.7	+65.4	+101.5
Silicon, p-type.....	7.8	+6.6	−1.1	+138.1	+93.6	+175
Silicon, n-type.....	11.7	−102.2	+53.4	−13.6	−7.6	−133

in semiconductor gages for applications around 25°C and possibly at elevated temperatures. We also notice that the first three materials in the table possess rather small π_{11} and π_{12} in contrast to π_{44}. For n-type silicon the situation is, however, reversed. The coefficient for hydrostatic loading, $\pi_{11} + 2\pi_{12}$ according to (6), is rather negligible in all of these semiconductors. We finally observe that all entries in the table are for bulk materials; i.e., they do not account for dimensional changes of a transducer due to applied strain. In the following, we sketch the explanation of these phenomena by solid state theory.

2.2 *Physical Explanation*

Fig. 1 illustrates some of the simplest ideas concerning the lattice structures of crystals. Although modern theories combine the concepts of statistics and quantum mechanics to represent the motion and energy content of a particle by wave functions, it will be adequate for a plausible description of piezoresistance to think of electrons in a corpuscular fashion. Fig. 1(a) illustrates the process by which "holes" and free electrons are generated or cancelled in a lattice structure by separation of an electron from its regular location in the structure of lattice bonds or by its "falling into" some "open" bond. In the lower left part of the figure we indicate the energy of the freely moving electron as that of the

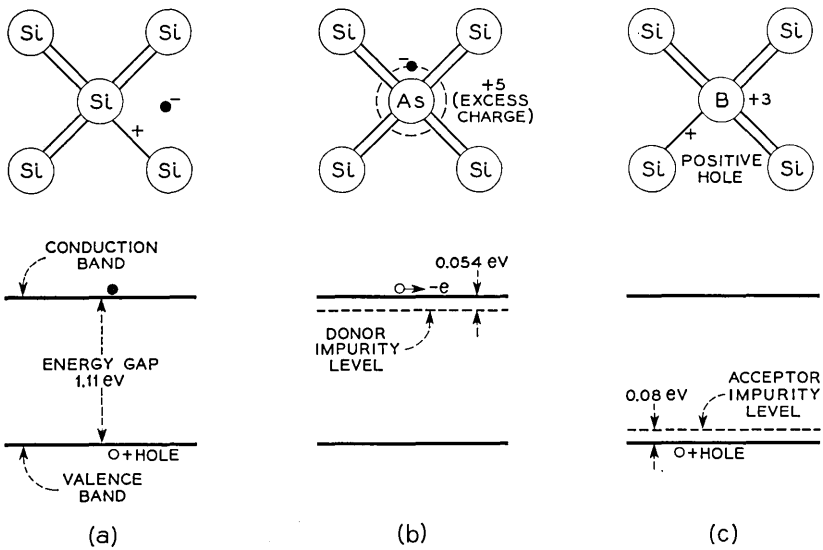


Fig. 1 — Schematic diagram of solid state lattice.

conduction band and the energy level in a closed bond as the *valence band*. The differential energy (1.11 ev for a silicon lattice) is liberated in the cancellation or absorbed in the generation of an open bond. Fig. 1(b) illustrates the state of affairs in the presence of a "donor" impurity in the lattice whose extra charge (one electron for arsenic) lacks relatively little energy (0.054 ev) to be raised into the conduction band and become a free electron. Conversely, the presence of an "acceptor" impurity offers an open-bond level which could be reached by some valence electron at the absorption of very little energy (0.08 ev for boron as acceptor), thus leaving a hole in its original location. These two processes constitute the basic transport phenomena in semiconductors.

Consider now in more detail the energy state of an electron in or above the conduction band. Quantum mechanics permits us to associate separate wave numbers $k_{1,2,3}$ with the components of its motion in the directions 1, 2, 3. In some media (e.g. a silicon lattice) it is possible for an electron to achieve a minimum energy which it requires to remain in the conduction band by several combinations of k_1, k_2 and k_3. These combinations are referred to as *band edge points*, since they constitute lower bounds for the energy required of a free electron. Fig. 2 shows such points in "k-space" for n-type silicon, where this space is resolved into components corresponding to the directions of crystal axes. An electron

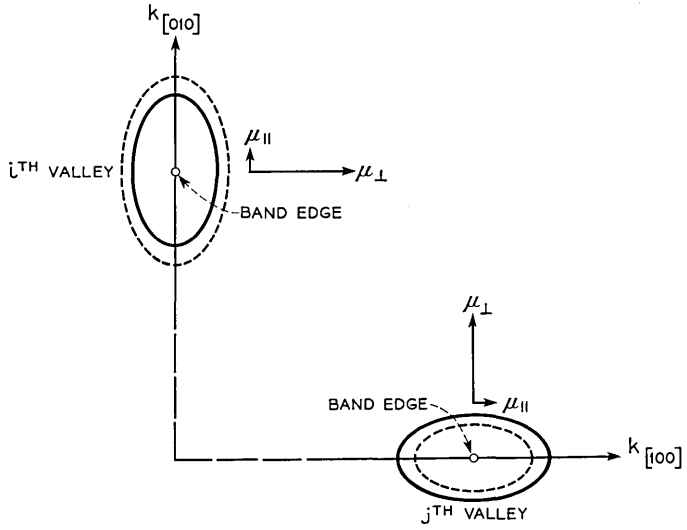


Fig. 2 — Constant energy surfaces in k-space.

with slightly more energy than is required at a band edge point may possess such energy by a variety of combinations in $k_{1,2,3}$ that describe a constant-energy surface about the band edge point. A family of such surfaces, centered on a band edge point, describes a so-called *energy valley* in k-space. In the case of silicon and germanium, these families consist of prolate ellipsoids of revolution that are aligned with the crystal axes. Since there are several band edge points, we speak of a *multivalley model*. The fact that the constant energy surfaces possess principal axes of unequal lengths may be interpreted to mean that the components of effective mass and mobility, $\mu_{1,2,3}$, of an electron or the components of conductivity, $\sigma_{1,2,3}$, in such a valley are different in the three principal directions. (In Fig. 2 we indicate mobility parallel to a crystal axes by $\mu_{\parallel}$ and transverse to the axis by $\mu_{\perp}$.) Consequently, these electrons make anisotropic contributions to the total conductivity of the lattice. If, however, all ellipsoids have the same proportions and all valleys are equally populated with electrons, the over-all conductivity of the lattice will be isotropic, as we have already observed for silicon and germanium in the unstressed state. The formulation of the above ideas in mathematical terms is known as *mobility theory*.* While this theory is advanced enough to account for the effects of impurity concentration, it also gives an adequate explanation of the variations of zero-stress resistivity with temperature (see Fig. 4 of Ref. 16). An experimental family of resistivity versus temperature characteristics for n-type silicon with various concentrations of phosphorus impurities is given in Fig. 3.

The theoretical treatment of piezoresistance now proceeds from the ideas of the multivalley model. It can be shown that the application of an anisotropic stress condition changes the relative energies, and hence changes the populations of these valleys. Thus, if a lattice started out as an isotropic conductor due to equal populations in the valleys at zero stress, an anisotropy is impressed on the total conductivity in the stressed state by the valleys which now attract the majority of electrons. Again, these ideas are treated in full detail in Ref. 16 (pp. 251–258 and Appendix C) and the resulting expressions for π_{st} yield relative magnitudes that are in essential agreement with Table I. Since the present theory

* Our lattice model serving as a basis for this theory is really a so-called simple multivalley model, in contrast to so-called degenerate single and multivalley models where the families of constant energy surfaces possess branch points. The relaxation time, τ, connected with electron mobility is only assumed to be a function of energy so that Maxwellian statistics are applicable. This simplification is adequate for a treatment of intra- and intervalley scattering due to acoustic excitation and neutral impurities, but not for cases involving highly anisotropic effective masses and subject to ionized-impurity scattering. A full treatment of this theory is given in Refs. 16 and 17.

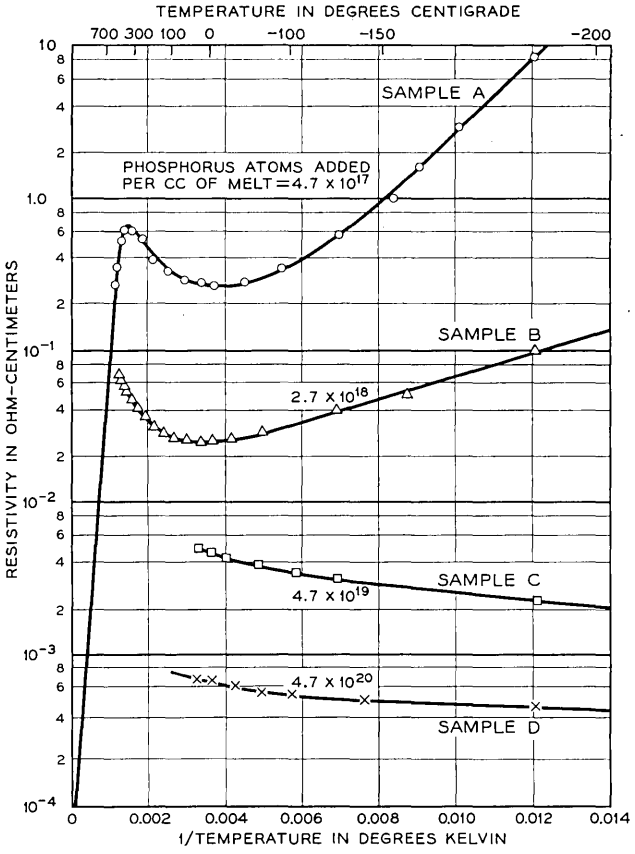


Fig. 3 — Resistivity vs. temperature for n-type silicon (from Ref. 27).

neglects certain minor effects of strain on the conductivity, it predicts that dilations along [100] directions will produce no piezoresistive effect in lattices with valleys on [111] axes since by symmetry the effect of the strain is the same on each valley. By the same token, no aggregative effect will be observed from strain in the [111] direction on material with valleys in the [100] direction. In the former case $\pi_{11} - \pi_{12} = 0$ and in the latter $\pi_{44} = 0$. Allowing for the fact that, in the absence of simplifications, no exact zeros will occur among piezoresistive coefficients, we observe from the experimental data in Table I that n-type germanium, for example, belongs to the first class, while n-type silicon is of the second kind. The theory also indicates a temperature dependence of π_{st} according to

$1/T$, which is experimentally confirmed for π_{11} in n-type silicon by Fig. 4. (Note that χ has been used only in this figure to indicate stress in place of T' or T_{kl} in order to avoid confusion with temperature, T. The reciprocal temperature dependence is quite exact in the temperature ranges where either intravalley or intervalley scattering predominate. Between these, the dependence is more complicated.)

III. GENERAL APPLICATIONS OF SEMICONDUCTOR STRAIN TRANSDUCERS

In the development of piezoresistive transducers, numerous embodiments other than strain gages had been worked on before the latter were being considered as self-contained units, to be applicable under a

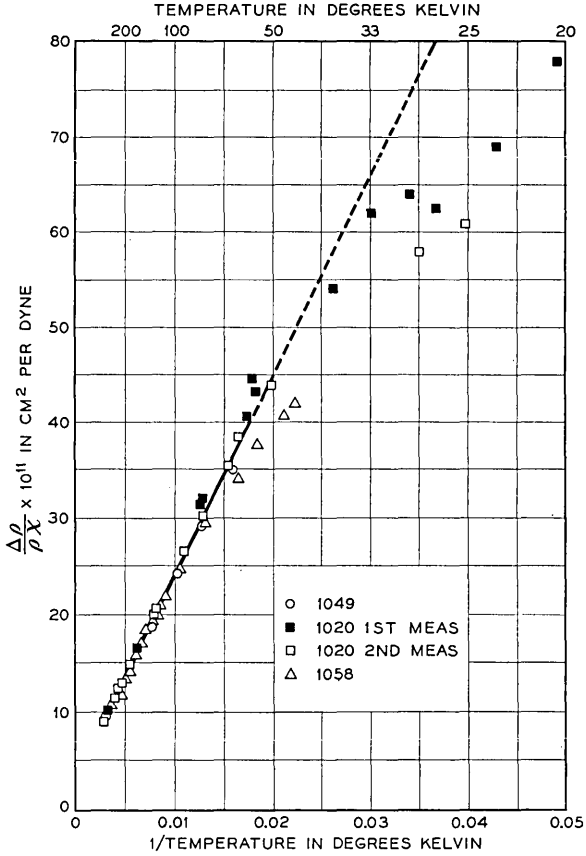


Fig. 4 — Piezoresistive sensitivity vs. temperature for n-type silicon (from Ref. 13).

great variety of test conditions. We shall, therefore, treat these earlier applications in the present section.

To the authors' knowledge, the first use of piezoresistive semiconductor elements for measuring strain was made by Bateman and McSkimin[18] at the suggestion of W. P. Mason, when they constructed a so-called bimorph cantilever unit, consisting of two bonded semiconductor lamellas, to measure small forces acting on the end of this transducer and subjecting it to flexure. By that time it was known that p-type silicon, and p- and n-type germanium exhibit large gage factors in the [111] direction, while n-type silicon is sensitive in the [100] direction, as was shown in Section II. If the circuitry surrounding a transducer is designed to provide a constant bias current, then the voltage change across the element will be the piezoresistive signal and can be expressed as

$$\Delta V = IRG\epsilon, \tag{7}$$

where

I = bias current,
R = nominal resistance of the transducer,
G = gage factor,
ϵ = mechanical strain.

Burns[19] subsequently used bimorph units and elements under axial stress in the construction of microphones. With a bimorph unit, the circuitry of Fig. 5 was employed, where the two elements constitute a tension and a compression arm, respectively, for the bridge, thus dou-

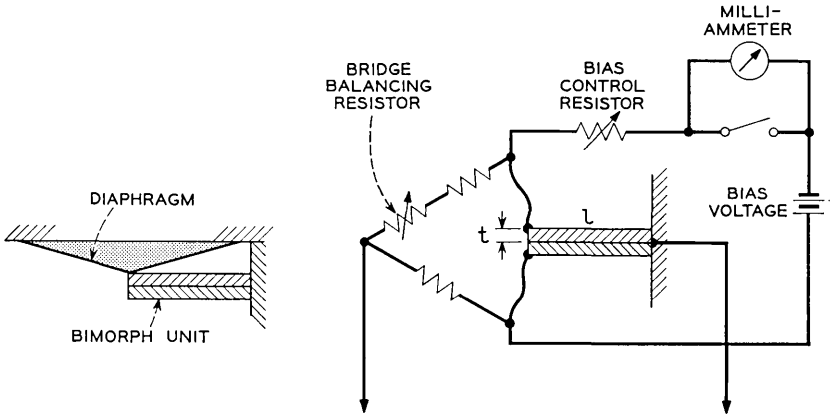


Fig. 5 — Electrical and mechanical arrangement for bimorph microphone (from Ref. 19).

bling the total signal. One pair of diagonal points of the bridge loop was controlled by the external bias circuit, the other served as signal output terminals. In this experimental configuration, the bimorph unit was made to support with one end the apex of a conical diaphragm and was suitably clamped at the other. Electrical connections were made by rhodium-plating the ends of the semiconductors and welding gold wires to them. Armstrong A-1 cement served as an insulating and bonding agent between the two halves of the bimorph unit. In the case of axially loaded elements, the technological details were quite similar, but with the transducer directed normally toward the center of the diaphragm. The dimensioning of both microphone configurations was designed such that the first natural frequency of the instrument was adequately high for practical applications. The merits of a piezoresistive rod microphone lie in the fact that it uses only about one-fiftieth of the power and one-tenth the current required to operate a carbon microphone for the same ac power output from a given acoustic power input.

Further uses of axially loaded and cantilver elements as dynamometers and dilatometers were suggested by Mason and Thurston.[20] In addition, a torsional transducer was developed, as illustrated in Fig. 6(a). It consisted of a cylindrical rod, built in at the left end and subjected to a torque at the right so that T_{yz} was positive on the front ($x > 0$, $z = 0$) and negative on the back ($x < 0$, $z = 0$). If now the transducer is made out of n-type germanium with [100] along the y-axis, and bias currents i_y are transmitted from the electrode 3 to 5 and 6 to 4, then the large piezoresistive shear coefficient π_{44} of this material will produce positive E_z along the front and back of the germanium cylinder; i.e., $E_z = \rho\pi_{44} i_y T_{yz}$ by an adaptation of the third equation in (4). This

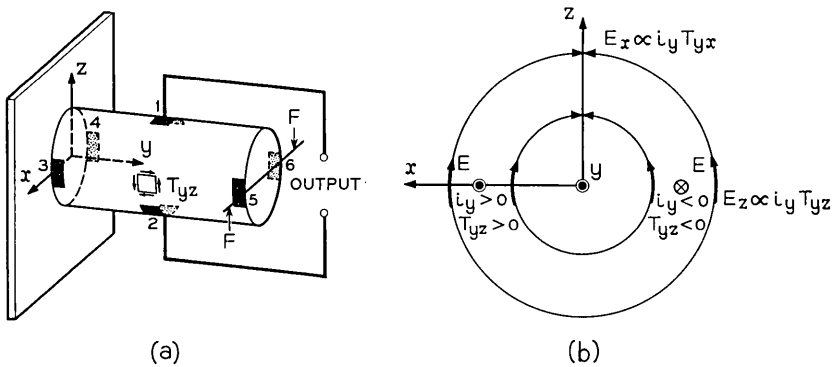


Fig. 6 — Schematic diagram of torsional transducer (from Ref. 20).

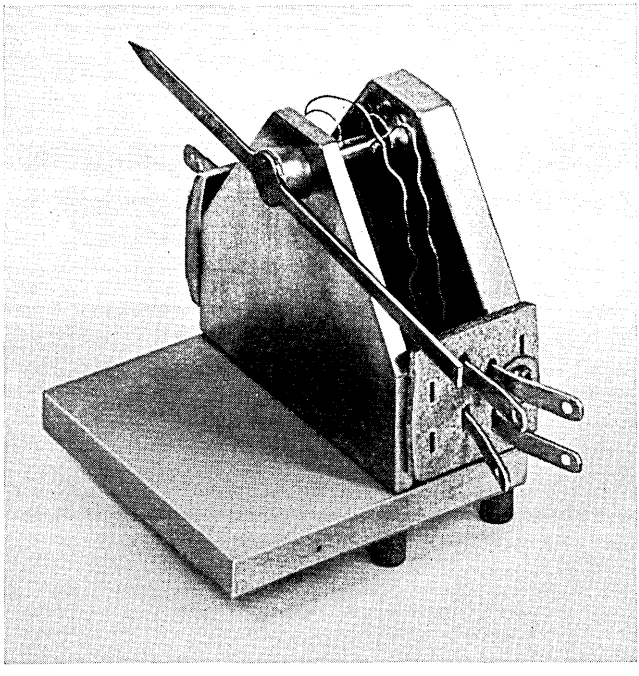


Fig. 7 — Experimental embodiment of torsional transducer (from Ref. 20).

voltage can be read as an output signal between terminals 1 and 2. The spatial relation between i_y , T_{yz} and E_z is illustrated further in Fig. 6(b).

An experimental transducer of this kind is shown in Fig. 7. The right end of the germanium rod is fastened into the upright end plate, while the left end is free to rotate under the torque produced by weights that are hung from the crossarm. Calibration curves for this device under different bias currents are shown in Fig. 8. We note that these data include the effect of cantilever bending in the rod, since the experimental setup subjected the germanium rod to a transverse force in addition to the torque. We also observe that its sensitivity might be improved if the rod were made bimorph with an insulating bonded interface in the yz-plane, which is bridged only by the electrodes 1 and 2. Thus, the parasitic currents from 3 to 4 and 6 to 5 would be eliminated.

Since the work conducted during the thirties on piezoresistive effects under hydrostatic pressure, not much attention had been given to this state of stress as a possible signal generator. In 1956, however, Flaschen, Sauer and Potter[21] discovered that lanthanum-doped barium strontium

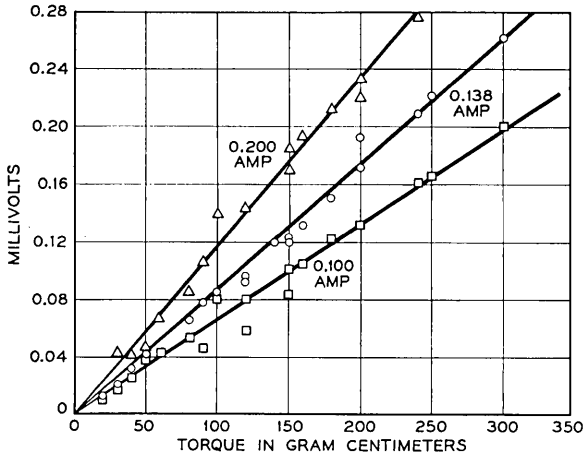


Fig. 8 — Characteristics of torsional transducer (from Ref. 20).

titanate showed rather high piezoresistive sensitivities. Luer[22] investigated its suitability as a hydrostatic pressure pickup, using material with a positive ion ratio of 0.600 barium, 0.397 strontium, 0.003 lanthanum. A pellet-shaped piece of $BaSrLaTiO_2$ was wetted with a layer of indium-gallium amalgam on its top and bottom surfaces and copper terminals were immersed into the amalgam. As an alternative terminating technique, nickel plating could be used to produce a surface to which wire terminals were bonded. Exploratory tests immediately indicated that this transducer was not only pressure sensitive but also showed strong response to temperature changes. Using an ice bath and other precautions, the isothermal calibration curve of Fig. 9 was obtained, which indicates that $\Delta R/R$ per 1000 psi was 2.5 per cent. Thus, the device was about 50 times as sensitive as commercially available load cells. It remains to be seen, however, whether convenient techniques for temperature control or compensation can be developed to make this a versatile miniature pressure gage.

An embodiment of piezoresistive elements for use as an accelerometer has been developed by Courtney-Pratt and Mason.[23] Fig. 10 shows the accelerometer frame with a seismic mass suspended in its center by a pair of sensing elements that are stressed axially in the y-direction. Additional pairs may be added in the x- and z-directions to form a tri-axial transducer.

The use of very small piezoresistive rods, similar to naturally grown whiskers, in dynamometers and extensometers has been pursued further by Forst and Geyling.[24] Fig. 11 shows the way in which germanium

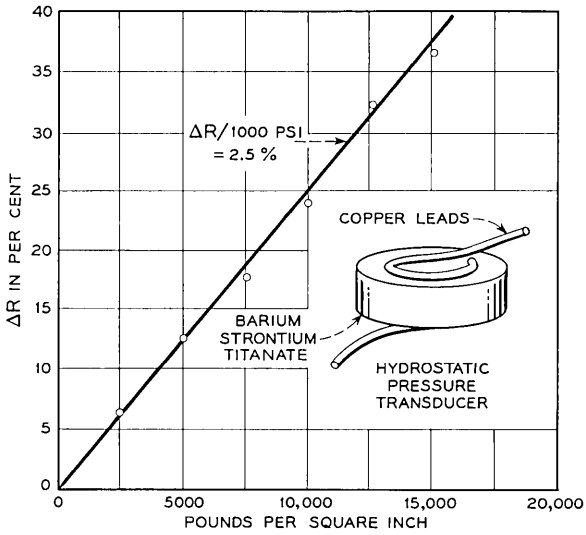


Fig. 9 — Schematic diagram and calibration curve of hydrostatic transducer.

rods with a 0.006-inch-square cross section were cut from an ingot in the [111] direction. After a 0.007-inch slab had been cut from the ingot, this was given a series of notches in the [111] direction and to a depth of 0.006 inch [Fig. 11(c)], whereupon the opposite side of the slab was lapped down until the material separated into the square rods [Fig. 11(d)].* Several ways of attaching gold lead wires to the ends of such

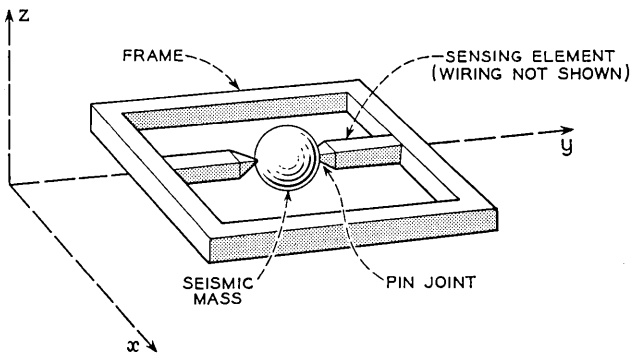


Fig. 10 — Schematic diagram of accelerometer.

* With preliminary equipment it was found advisable to use a slab thickness of 0.015 inch and notch depth of 0.012 inch. The final dimensions of 0.006 × 0.006 inch were then obtained by lapping *both* sides of the work piece in Fig. 11(d).

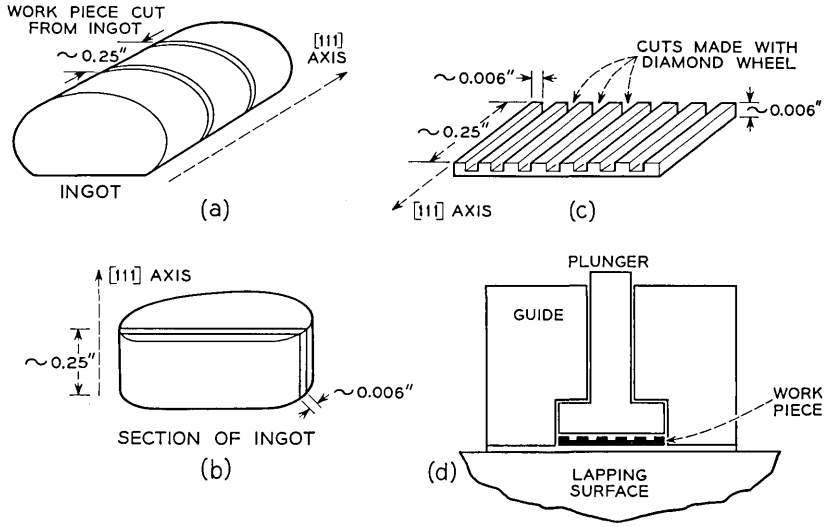


Fig. 11 — Method of cutting germanium rods.

rods were attempted, including such techniques as gold-plating the ends of the rods and then soldering to them, as well as sintered bonds and thermocompression bonds. None of these are fundamentally impossible, and they could be developed into a routine technology. The easiest for laboratory work, however, turned out to be a simple fusion process, which was carried out somewhat above the germanium-gold eutectic temperature. A rod was placed on a piece of flat stock with a pair of gold wires stretched taut across its ends or held down by a retaining spring (Fig. 12) to insure good mechanical contact and to hold the rod in place. Upon heating to 410°C in a nitrogen atmosphere, flow occurred between the germanium and the gold, producing the terminations shown in Fig. 13. Due to a 0.1 per cent doping of antimony in the gold wire,

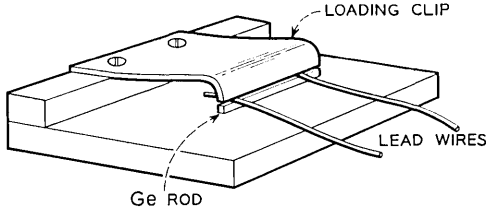


Fig. 12 — Method of fusing gold to germanium.

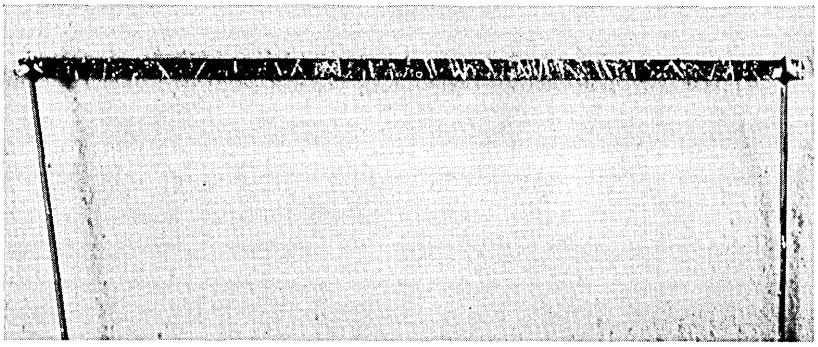


Fig. 13 — Experimental germanium rod and terminations.

these terminations were purely ohmic. As is evident from this discussion, more can be done to mechanize the fabrication of these basic elements and to increase the yield in matched pairs which act as "pickup" and "compensator." Fig. 13 also shows that refined cutting and etching techniques should produce rods with a better surface finish and possibly fewer dislocations, so that the maximum tensile strain in these elements could be increased considerably. At present, the basic techniques are being extended to the use of silicon rods.

The n-type germanium that provided the first experimental elements was manufactured with a zero-stress resistivity of 0.03 ohm-cm, so that a rod of 0.006 × 0.006-inch cross section and 0.25-inch length had a basic resistance of approximately 100 ohms. This was arranged for convenient use of the SR-4 indicator with these elements. Another advantage of this low resistivity lies in the fact that semiconductors with low ρ show less internal contact noise in low-frequency applications than does material with high ρ.

Fig. 14 shows an application of such rods in a tensile extensometer. A tension specimen, h, is held between the grips, g, which transfer the differential motion between gage points on the specimen, i.e., the tensile strain, through the sliding bars f and e to a cantilever-flexing bar of spring steel, k. The structure a,b,c serves to hold the ball bushings, d, which guide the bars f and e. Safety collars, i and i′, on the sliding bars guard against specimen failure. The steel spring, k, carries longitudinal germanium rods on its top and bottom face. These are attached with Araldite CN502 room curing cement, which also serves as insulator. The wiring post, m, supports the lead wires, and the germanium rods are connected into the SR-4 indicator as "active" and "compensating" transducers, to produce twice the signal of a single element.

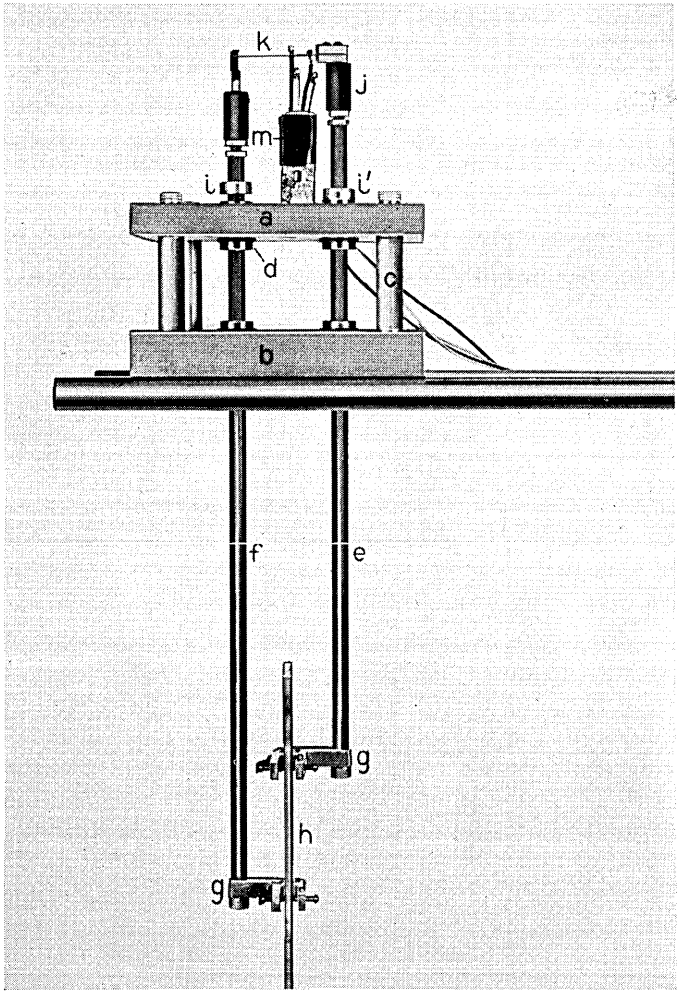


Fig. 14 — Piezoresistive extensometer.

The first experiments immediately showed that these highly sensitive transducers introduced considerable noise into the signal, due to environmental vibrations, draft and varying thermal conditions. Since the present experiment did not call for all the sensitivity available, each element was partially shunted out to reduce its sensitivity to 10 times that of a wire or foil gage and make the noise negligible. Fig. 15 shows typical calibration curves for this extensometer, which were run by controlling

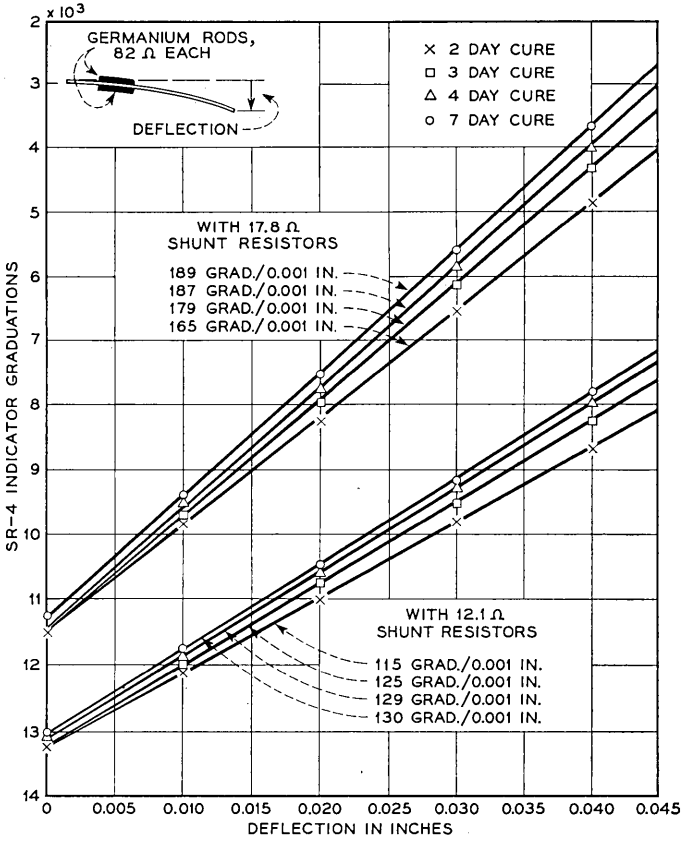


Fig. 15 — Calibration of extensometer.

the relative motion of the gage points, g, with a micrometer and using the arbitrary value of $G = 2$ on the gage factor dial of the indicator. Sensitivities of 189 indicator divisions per mil of extensometer movement and 130 divisions per mil compared favorably with the 14 divisions per mil obtained when A-7 wire gages were used in place of the germanium rods. The shakedown observed in the curves and the fact that some loss of sensitivity occurred with rising ambient temperature indicate that creep and elastic deformation in the bond of the germanium rods is a problem to be kept in mind.[24]

In cases where extremely rigid devices are needed for the measurement of small tensile or compressive forces, the semiconductor elements under axial load are, at present, the only transducers sensitive enough

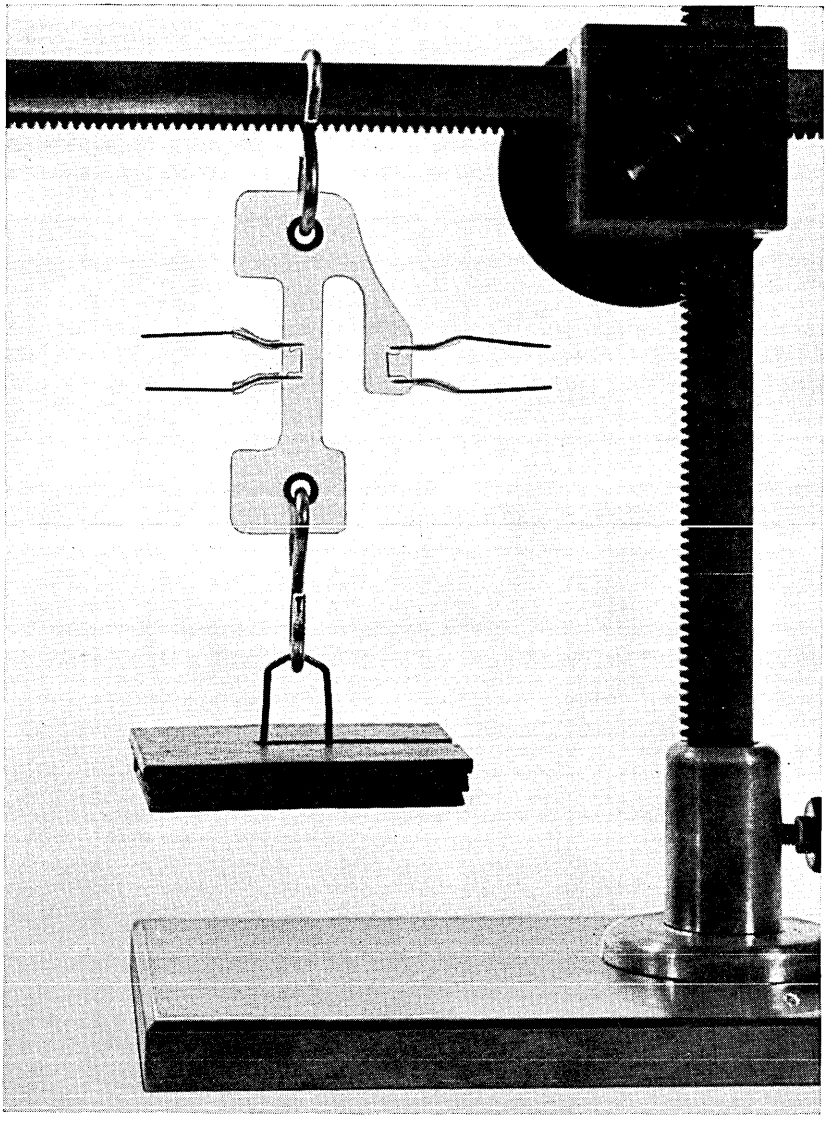


Fig. 16 — Tension dynamometer.

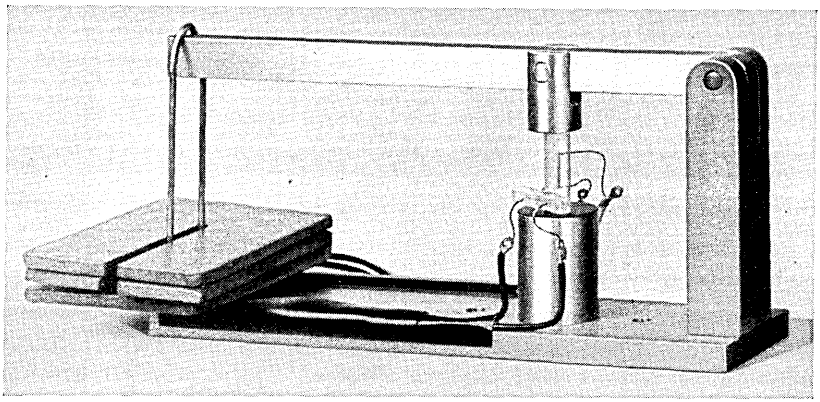


Fig. 17 — Compression dynamometer.

for this purpose. Figs. 16 and 17 show a tension and a compression dynamometer each with an active element in line with the load and the compensator in a monolithic extension of the encapsulation but designed to be strain-free. Typical sensitivities of these transducers were 544 and 637 divisions per pound (with $G = 2.00$ on the indicator). Of course, these sensitivities may be increased further by decreasing the active cross section of the encapsulation.

In closing this section, we note that nearly all the embodiments that were mentioned exhibit negligible deformations in the course of generating a signal. This may be a decisive factor in their application to dynamic problems, where resonances and self-sustained vibrations in the instrumentation must be avoided.

IV. STRAIN GAGE APPLICATIONS

After most of the devices in Section III had been developed,[25] interest shifted to strain gage applications of semiconductor elements. The two main problems in using semiconductor rods for this purpose are: (a) the mechanical one of making a piezoresistive element follow the surface strain of the specimen it is attached to and (b) the task of accurately determining the temperature response of the transducer and compensating for it.

4.1 *Bond Rigidity of Strain Gages*

In order to make a sensing element represent the true deformation of a test model, one must first of all insure that the semiconductor material

is capable of undergoing the strains to be measured. Indeed, germanium and silicon, if machined carelessly, will show little resilience and a rather low ultimate tensile strain. We have already remarked, however, that the maximum allowable strain may be increased substantially by proper surface treatment. Some information on proper machining techniques for semiconductors is contained in Ref. 26, and further work along these lines is in progress.

The second problematic aspect of mechanical gage performance is the adhesion of sensing elements to the model surface. We noted in Section III that some tests (Fig. 15) gave evidence of deformation in the gage cement. This is hardly surprising if resins are used for the bond and one may expect a noticeable increase in creep with rising temperature. For so-called high-temperature applications some special bonding agent will have to be used.

The currently most popular high-temperature cement for strain gages is Allen PBX. It requires a curing cycle with soaking periods at 200°F and 600°F. Since germanium has a thermal coefficient of expansion of $\cong 6 \times 10^{-6}$ per °C and most structural steels have $\alpha \cong 12 \times 10^{-6}$ per °C, some special precautions will be necessary in mounting semiconductor gages on ferrous specimens. Laboratory tests have so far been restricted to Kovar test models where $\alpha = 5 \times 10^{-6}$ per °C, and hence there is little chance of gage failure during the curing cycle. A typical Kovar dumbbell specimen is shown in Fig. 18, and its calibration curve, using p-type silicon rods at room temperature, is given in Fig. 19. The same illustration also lists the sensitivities of these gages at several elevated temperatures.

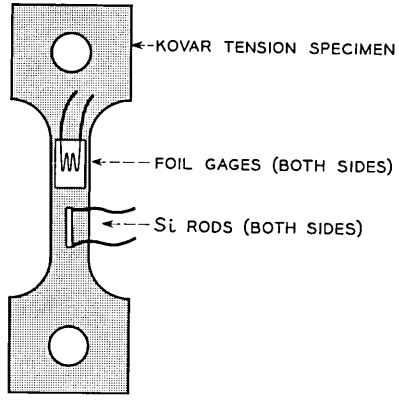


Fig. 18 — Schematic diagram of test specimen for strain gage applications.

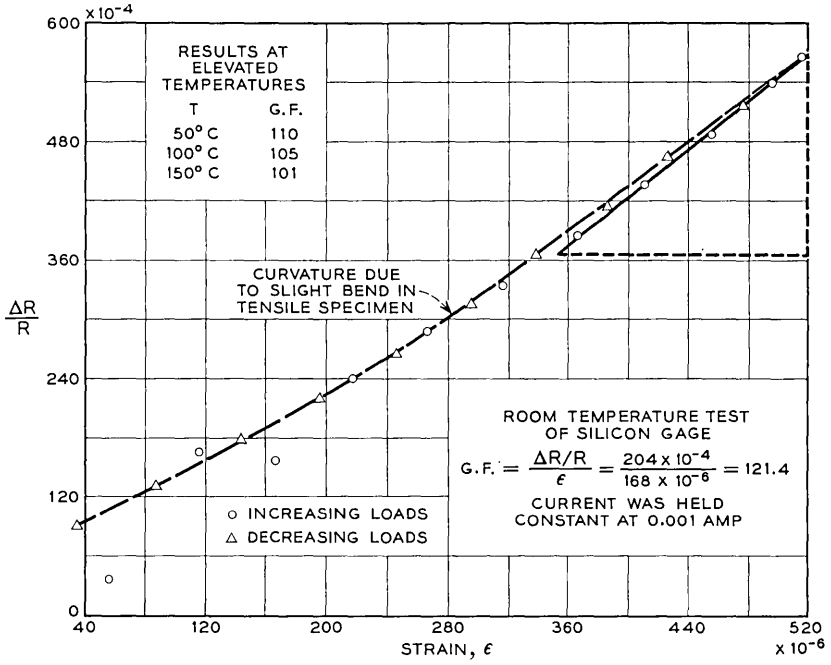


Fig. 19 — Calibration of test specimen for strain gage applications.

One way of observing creep in the bond consists of making measurements of gage signal against time at constant strain in the dumbbell. As an alternative, one may mount a commercial high-temperature gage on the same specimen and, knowing its gage factor, find its strain from the SR-4 signal. If the gage creeps, this value should be less than the strain of the specimen computed from the load. Of course, there is little correlation between the creep of a foil or wire gage and that of a semiconductor rod, since the latter tends to have a smaller ratio of surface to cross-sectional area—even at sizes of 0.001 × 0.001 inch.* The most reliable, but also the most elaborate, instrumentation for measurements of bond rigidity would consist of placing a suitable mechanical gage (e.g., a Huggenberger tensometer) right on the semiconductor rod and comparing its readings with the computed specimen strain. Needless to say, slippage of a gage would make its apparent gage factor fall below the true one and produce a zero-load signal at the end of a test.

* Clearly, the small size of natural whiskers makes them attractive for the minimization of creep. Their twinning points will add to the bond if they are oriented in a way to permit a snug fit of the whisker against the model surface.

4.2 *Temperature Compensation for Strain Gages*

As we have noted in Section II, the zero-stress resistivity as well as the piezoresistive coefficients of a semiconductor change with temperature. If we are to use these transducers in an extended range around room temperature, not to mention regions like 200°–600°C as required by many applications, both phenomena require compensation.

Fig. 3 indicates the change of ρ with T for n-type silicon, which is the phenomenon one usually eliminates in existing practice by wiring a "dummy" gage into another branch of the bridge as a compensator. This "dummy" needs to be installed near the active gage, preferably on the test model itself, to attain the same ambient conditions, but must remain strain-free. Since these two conditions turn out to be antagonistic in most high-temperature applications, this method ceases to be a satisfactory solution. To aggravate the problem, the high strain sensitivity of semiconductors will magnify the signals from any differential thermal strains between the active and compensating elements over what it is for wire or foil gages.

Several courses of action can be taken in this situation:

i. Subminiature thermocouples are presently on the market that can be installed sufficiently close to each active gage to furnish a temperature signal along with the strain reading. These temperature data can be turned into a correction of the strain reading through the known ρ versus T characteristic of the semiconductor. This may be done either manually or by a compensating adjustment in the indicator circuitry. More elaborate instrumentation might even feature automatic control of this adjustment by a servo.

ii. If the ρ versus T characteristic is difficult to represent with sufficient accuracy in the automatic version of scheme i, we might return to the conventional setup involving a compensating gage. But this element might be contained in a separate furnace whose temperature is automatically steered to match the thermocouple signal from a given active gage.

iii. A third possibility consists of making compensating elements out of semiconductor material with the same ρ versus T characteristic as the active gages but cutting these compensators along the axis of minimum piezoresistive sensitivity ([111] for n-type silicon, and [100] for p-type silicon and p- and n-type germanium). Each gage installation would consist of a closely spaced parallel pair of elements (one "active" and one "dummy"), which are to be wired into opposite branches of the indicator. Thermal signals cancel in the usual way, since cubic lattices exhibit isotropic zero-stress resistivity (i.e., the same ρ on [111]

and [100]), but the greatly different piezoresistive coefficients of the two rods would still produce a large signal due to mechanical strain.

In order to compensate for temperature changes of piezoresistive sensitivity, we observe that the coefficients π_{st} change according to $1/T$. This was shown by Fig. 4 for π_{11} of n-type silicon, where T is in degrees Kelvin. Although the response of wire and foil gages has the same form, its factor of proportionality is considerably less than for semiconductors, the ratio being about one to ten. While the sensitivity decreases, its variability with temperature, $d\pi_{st}/dT$, also diminishes with T, which constitutes a slight consolation for the inventor of compensating schemes. No one system has as yet been envisioned to cover the entire operating range of semiconductor elements. In the following, we describe an arrangement suggested by Mason for compensation over the range of $-70°$ to $+100°$C.

In the bridge circuit of Fig. 20, let the impressed voltage be E_1 and each of the branch resistors be a strain gage of zero-stress resistance R_0, which is, of course, temperature-dependent. Let R_6 be large and very constant. Then the expression for E_0, the voltage drop across the output resistor R_5, may be obtained in the simplified form given in Fig. 20, where ΔR is the change in gage resistance due to mechanical strain. If only R_5 were subjected to temperature variations but all other circuit

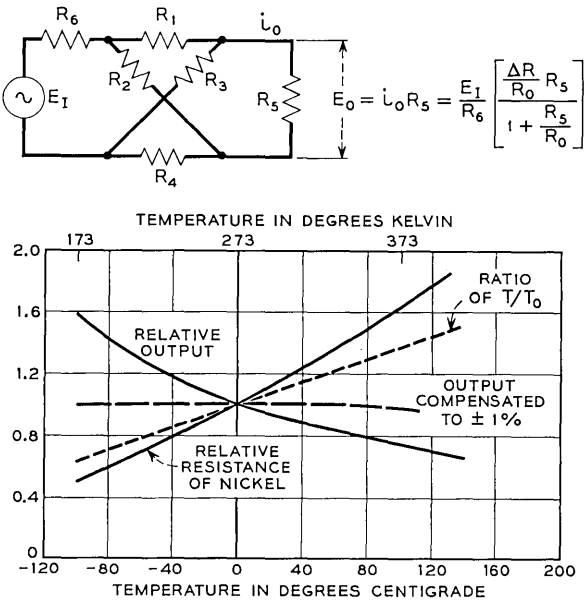


Fig. 20 — Proposed scheme for temperature compensation.

parameters were held fixed, its voltage output would be the ascending curve in Fig. 20. Conversely, thermal changes in $R_{1,2,3,4}$ would produce the descending curve for E_0. The formula indicates that proper choice of R_0/R_5 will provide a cancellation of these effects and will produce an output that is compensated within 1 per cent between $-70°$ and 100°C at a decrease in sensitivity by the factor 0.37. This may, for example, be accomplished with a nickel resistor for R_5.

V. CONCLUSIONS AND PROBLEMS FOR THE FUTURE

The foregoing discussions indicate that semiconductor materials show considerable promise for use in a variety of piezoresistive transducers. They are capable of handling both dc and ac signals and, due to a sensitivity about 70 times that of wire gages at room temperature, they eliminate the need for signal amplification in many situations. Their small dimensions permit subminiature embodiments for airborne applications and installations with severe space restrictions.

Some open questions regarding slippage and temperature compensation of these gages have already been pointed out in Section IV. To this list of current problems we add the need for adequate electrical filters to accompany the gages when their full sensitivity is to be utilized at minimum interference from noise. A final topic to be investigated is the resistance of such elements to mechanical shock and their fatigue characteristics under sustained vibrations.* Clearly, future efforts in the manufacture of dislocation-free semiconductor rods or the use of natural whiskers will have considerable bearing on this issue.

REFERENCES

1. Bridgman, P. W., Effect of Tension on the Resistance of Metals, Proc. Am. Acad. Arts Sci., **60,** 1925, p. 423.
2. Bridgman, P. W., The Effect of Homogeneous Mechanical Stress on the Electrical Resistance of Crystals, Phys. Rev., **42,** 1932, p. 858.
3. Allen, M., The Effect of Tension on the Electrical Resistance of Single Bismuth Crystals, Phys. Rev., **42,** 1932, p. 848.
4. Allen, M., The Effect of Tension on the Electrical Resistance of Single Antimony Crystals, Phys. Rev., **43,** 1933, p. 569.
5. Cookson, J. W., Theory of the Piezoresistive Effect, Phys. Rev., **47,** 1935, p. 194.
6. Allen, M., The Tension Coefficients of Resistance of Hexagonal Crystals Zinc and Cadmium, Phys. Rev., **49,** 1936, p. 248.
7. Allen, M., The Effect of Tension on the Electrical Resistance of Single Tetragonal Tin Crystals, Phys. Rev., **52,** 1937, p. 1246.
8. Lawrence, R., Temperature Dependence of Drift Mobility in Germanium, Phys. Rev., **89,** 1953, p. 1295.

* Some information on high-frequency fatigue endurance of silicon has been obtained by W. P. Mason with ultrasonic methods.

9. Smith, C. S., Piezoresistance Effect in Silicon and Germanium, Phys. Rev., **94,** 1954, p. 42.
10. Keyes, R. W., Temperature Dependence of the Elasto-resistance in N-Type Germanium, Phys. Rev., **100,** 1955, p. 1104.
11. Pearson, G. L., Read, W. T. and Feldmann, W. L., Deformation and Fracture of Small Silicon Crystals, Acta Metall., **5,** 1957, p. 181.
12. Bowers, R., Magnetic Susceptibility of Germanium, Phys. Rev., **108,** 1957, p. 683.
13. Morin, F. J., Geballe, T. H. and Herring, C., Temperature Dependence of Piezoresistance of High-Purity Silicon, Germanium, Phys. Rev., **105,** 1957, p. 525.
14. Pollack, M., Piezoresistance in Heavily Doped N-Type Germanium, Phys. Rev., **111,** 1958, p. 798.
15. Tuzzolino, A., Piezoresistance Constants of P-Type InSb, Phys. Rev., **109,** 1958, p. 1980.
16. Herring, C., Transport Properties of a Many-Valley Semiconductor, B.S.T.J., **34,** 1955, p. 237.
17. Herring, C. and Vogt, E., Transport and Deformation Potential Theory for Many-Valley Semiconductors with Anisotropic Scattering, Phys. Rev., **101,** 1956, p. 944.
18. Bateman, T. B. and McSkimin, H. T., private communication.
19. Burns, F. P., Piezoresistive Semiconductor Microphone, J. Acoust. Soc. Am., **29,** 1957, p. 248.
20. Mason, W. P., and Thurston, R. N., Piezoresistive Materials in Measuring Displacement, Force and Torque, J. Acoust. Soc. Am., **29,** 1957, p. 1096.
21. Sauer, H. A., Flaschen, S. S., and Hoesterey, T. C., Piezoresistance and Piezocapacitance Effects in Barium Strontium Titanate Ceramics, J. Am. Cer. Soc., **42,** 1959, p. 363.
22. Luer, H. L., private communication.
23. Courtney-Pratt, J. S. and Mason, W. P. private communication.
24. Forst, J. J. and Geyling, F. T., Applications of Semiconductor Transducers in Strain Gages and Rigid Dynamometers, Proc. Soc. Exp. Stress Anal., **17,** 1960, p. 143.
25. Mason, W. P., Semiconductors in Strain Gages, Bell Labs. Record, **37,** 1959, p. 7.
26. Dash, W. C., Growth of Silicon Crystals Free from Dislocations, J. App. Phys., **30,** 1959, p. 459.
27. Shockley, W., *Electrons and Holes in Semiconductors*, D. Van Nostrand Co., New York, 1950.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

ABEL, J. L., see Chynoweth, A. G.

ABRAHAMS, S. C., see Prince, E.

ALLEN, F. G., EISINGER, J., HAGSTRUM, H. D., and LAW, J. T.

Cleaning of Silicon Surfaces by Heating in High Vacuum, Monograph 3444.

ALLISON, H. W. and SAMELSON, H.

Diffusion of Aluminum, Magnesium, Silicon and Zirconium in Nickel, Monograph 3445.

ANDERSON, E. W., see MCCALL, D. W.

ANDERSON, P. W.

Spectral Diffusion, Phonons and Paramagnetic Spin-Lattice Relaxation, Monograph 3386.

ANDERSON, P. W.

New Approach to the Theory of Superexchange Interactions, Monograph 3415.

AUGUSTYNIAK, W. M., see Brown, W. L.

AUSTIN, T. L., FAGEN, R. E., PENNEY, W. F. and RIORDAN, J.

The Number of Components in Random Linear Graphs, Monograph 3446.

BAKER, A. N., GOLDEY, J. M., and ROSS, I. M.

Recovery Time of PNPN Diodes, Monograph 3403.

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

BANGERT, J.T.

Practical Applications of Time Domain Theory, Monograph 3404.

BATTERMAN, B. W.

X-Ray Measurement of the Atomic Scattering Factor of Iron, Monograph 3416.

BEMSKI, G.

Paramagnetic Resonance in Electron Irradiated Silicon, Monograph 3417.

BOWERS, K. D. and MIMS, W. B.

Paramagnetic Relaxation in Nickel Fluosilicate, Monograph 3418.

BRAILSFORD, A. D., see Galt, J. K.

BROUT, R.

Statistical Mechanical Theory of a Random Ferromagnetic System, Monograph 3447.

BROWN, W. L., AUGUSTYNIAK, W. M. and WAITE, T. R.

Annealing of Radiation Defects in Semiconductors, Monograph 3420.

BROWN, W. L. and AUGUSTYNIAK, W. M.

Defect Formation in Electron Irradiation of N-Type Germanium, Monograph 3419.

BUCHANAN, D. N. E., see Wertheim, G. K.

BUDENSTEIN, P. P., see Ohl, R. S.

BURBANK, R. D., see Heidenreich, R. D.

BURRUS, C. A., see Ohl, R. S.

CETLIN, B. B., see Galt, J. K.

CHYNOWETH, A. G. and ABEL, J. L.

Polarization Reversal in Ferroelectric Triglycine Sulfate, Monograph 3421.

COLLINS, R. J.

The Mechanism and Defect Responsible for Edge Emission in CdS, Monograph 3422.

COMPTON, V. B. and MATTHIAS, B. T.

Laves Phase Compounds of Rare Earths and of Hafnium with Noble Metals, Monograph 3423.

DAVID, E. E., JR., MATHEWS, M. V. and MC DONALD, H. S.

Experiments with Speech Using Digital Computer Simulation, Monograph 3405.

DAVIS, J. L., DEVLIN, G. E., JACCARINO, V. and SCHAWLOW, A. L.

Structure-Sensitivity of High-Frequency NMR in Powdered Antiferromagnetic MnF_2, Monograph 3424.

DEVLIN, G. E., see Davis, J. L.

DOUGLASS, D. C., see McCall, D. W.

EISINGER, J. and LAW, J. T.

Adsorption of Oxygen on Silicon, Monograph 3318.

EISINGER, J., see Allen, F. G.

FAGEN, R. E., see Austin, T. L.

FEHER, G. and GERE, E. A.

Electron Spin Resonance Experiments on Donors in Silicon, Monograph 3388.

FLANAGAN, J. L.

Resonance-Vocoder and Baseband Complement: Hybrid Speech Transmission, Monograph 3407.

FLASCHEN, S. S., see Sauer, H. A.

FLÜGGE, W. and ZAJAC, E. E.

Bending Impact Waves in Beams, Monograph 3425.

FRISCH, H. L. and LUNDBERG, J. L.

Viscometric Criterion of Polymer Polydispersity, Monograph 3408.

FRISCH, H. L., see Reiss, H.

GALT, J. K., YAGER, W. A., MERRITT, F. R., CETLIN, B. B. and BRAILSFORD, A. D.

Cyclotron Absorption in Metallic Bismuth and Its Alloys, Monograph 3390.

GEBALLE, T. H.

Radiation Effects in Semiconductors, Monograph 3409.

GELLER, S., see Gilleo, M. A.

GELLER, S., see Wernick, J. H.

GERE, E. A., see Feher, G.

GIBBONS, D. F. and RENTON, C. A.

Change in Velocity of Sound between Normal and Superconducting States in Tin, Monograph 3389.

GIBBONS, D. F.

Thermal Expansion Coefficients of Manganese Fluoride, Monograph 3451.

GILLEO, M. A. and MITCHELL, D. W.

Ferrimagnetism and Crystal Chemistry of Substituted Manganese-Tin Spinels, Monograph 3428.

GILLEO, M. A. and GELLER, S.

Interaction of Magnetic Ions in Gd_3, Mn_2, Ge_2, GaO_{12} and Related Garnets, Monograph 3427.

GOLDEY, J, M., see Baker, A. N.

GYORGY, E. M. and HAGEDORN, F. B.,

Uniform Rotational Flux Reversal of Ferrite Toroids, Monograph 3452.

HAGEDORN, F. B., see Gyorgy, E. M.

HAGSTRUM, H. D., see Allen, F. G.

HEIDENREICH, R. D., NESBITT, E. A. and BURBANK, R. D.

Magnetic Annealing in Perminvar, Monograph 3429.

HIGHLEYMAN, W. H. and KAMENTSKY, L. A.

Generalized Scanner for Pattern and Character Studies; Picture Processing, Monograph 3410.

HOESTEREY, D. C., see Sauer, H. A.

HOGG, D. C.

Effective Antenna Temperatures Due to Oxygen and Water Vapor in the Atmosphere, Monograph 3453.

HOPFIELD, J. J.

A Theory of Edge-Emission Phenomena in CdS, ZnS and ZnO, Monograph 3430.

INGRAM, S. B.

Education for the Age of Technology, Monograph 3454.

JACCARINO, V., see Davis, J. L.

KAMENTSKY, L. A., see Highleyman, W. H.

KINOSHITA, T.

Ground State of the Helium Atom. II, Monograph 3423.

KRUSEMEYER, H. J.

Surface Conductivity of ZnO Crystals, Monograph 3385.

LAW, J. T., see Allen, F. G.

LAW, J. T., see Eisinger, J.

LEBOWITZ, J. L., see Reiss, H.

LI, TINGYE

Study of Spherical Reflectors as Wide-Angle Scanning Antennas, Monograph 3411.

LLOYD, S. P.

A Sampling Theorem for Stationary (Wide Sense) Stochastic Processes, Monograph 3433.

LOUISELL, W. H.

A Three-Frequency Electron Beam Parametric Amplifier and Frequency Converter, Monograph 3346.

LUNDBERG, J. L., see Frisch, H. L.

MARTENS, H. H.

Two Notes on Machine Learning, Monograph 3457.

MATHEWS, M. V., see David, E. E., Jr.

MATTHIAS, B. T., see Compton, V. B.

MATTHIAS, B. T., see Suhl, H.

McCALL, D. W., DOUGLASS, D. C. and ANDERSON, E. W.

Diffusion in Ethylene Polymers, IV., Monograph 3325.

MC DONALD, H. S., see David, E. E., Jr.

MERRITT, F. R., see Galt, J. K.

MILLER, R. C. and SAVAGE, A.

Further Experiments on the Sidewise Motion of 180° Domain Walls in $BaTiO_3$, Monograph 3459.

MIMS, W. B., see Bowers, K. D.

MITCHELL, D. W., see Gilleo, M. A

NESBITT, E. A., see Heidenreich, R. D.

OHL, R. S., BUDENSTEIN, P. P., and BURRUS, C. A.

Improved Diode for the Harmonic Generation of Millimeter and Submillimeter Waves, Monograph 3460.

PAULL, M. C. and UNGER, S. H.

Minimizing the Number of States in Sequential Switching Functions, Monograph 3463.

PENNEY, W. F., see Austin, T. L.

PIERCE, J. R.

Conservation Principles in Operation of Wave-Type Parametric Amplifiers, Monograph 3464.

PREZIOSI, S., see Van Uitert, L. G.

PRINCE, E. and ABRAHAMS, S. C.

Single-Crystal Automatic Neutron Diffractometer, Monograph 3435.

RASPANTI, M.

SPUD, A Stored-Program Universal Demonstrator for Computer Training, Monograph 3493.

RENTON, C. A., see Gibbons, D. F.

REISS, H., FRISCH, H. L. and LEBOWITZ, J. L.

Statistical Mechanics of Rigid Spheres, Monograph 3466.

RIORDAN, J., see Austin, T. L.

ROBERTS, S. W.

Control Chart Tests Based on Geometric Moving Averages, Monograph 3436.

ROSS, I. M., see Baker, A. N.

RUTHROFF, C. L.

Some Broadband Transformers, Monograph 3437.

SAMELSON, H., see Allison, H. W.

SAUER, H. A., FLASCHEN, S. S. and HOESTEREY, D. C.

Piezoresistance and Piezocapacitance in Barium Strontium Titanate Ceramics, Monograph 3412.

SAVAGE, A., see Miller, R. C.

SCHAWLOW, A. L., see Davis, J. L.

SCHROEDER, M. R.

Measurement of Sound Diffusion in Reverberation Chambers, Monograph 3469.

SCHWENZFEGER, E. E.

A High-Volume, High-Speed Weather Information Distribution System, Monograph 3496.

SIMONE, C. F.

Equivalent Ladder Networks by the Use of Signal Flow Graphs, Monograph 3438.

SINDEN, F. W.

Mechanisms for Linear Programs, Monograph 3497.

SUHL, H. and MATTHIAS, B. T.

Impurity Scattering in Superconductors, Monograph 3387.

SWANEKAMP, F. W., see Van Unitert, L. G.

THEUERER, H. C.

Purification of $SiCl_4$ by Adsorption Techniques, Monograph 3471.

TURNER, D. R.

Electroplating Metal Contacts on Germanium and Silicon, Monograph 3413.

TURNER, D. R.

Junction Delineation on Silicon in Electrochemical Displacement Plating Solutions, Monograph 3439.

UNGER, S. H.

Hazards and Delays in Asynchronous Sequential Switching Circuits, Monograph 3414.

UNGER, S. H., see Paull, M. C.

VAN UITERT, L. G., SWANEKAMP, F. W. and PREZIOSI, S.

Ferrimagnetic Resonance and Phase Distributions in Sintered Yttrium Iron Garnets, Monograph 3473.

WAITE, T. R., see Brown, W. L.

WEISS, J. A.

A Phenomenological Theory of the Reggia-Spencer Phase Shifter, Monograph 3355.

WERNICK, J. H. and GELLER, S.

Transition Element-Rare-Earth Compounds with the Cu_5Ca Structure, Monograph 3474.

WERTHEIM, G. K.

Recombination Properties of Bombardment Defects in Semiconductors, Monograph 3440.

WERTHEIM, G. K.

Recombination Properties of Nickel in Germanium, Monograph 3441.

WERTHEIM, G. K.

Temperature-Dependent Defect Production in Bombardment of Semiconductors, Monograph 3442.

WERTHEIM, G. K. and BUCHANAN, D. N. E.

Electron-Bombardment Damage in Oxygen-Free Silicon, Monograph 3443.

YAGER, W. A., see Galt, J. K.

ZAJAC, E. E., see Flügge, W.

Contributors to This Issue

A. A. ALEXANDER, B.S., 1943, Newark College of Engineering; M.S. in E.E., 1949, Stevens Institute of Technology; M.B.A., 1955, New York University; New Jersey Bell Telephone Co., 1947–56; American Telephone and Telegraph Co., 1956—. While with New Jersey Bell, Mr. Alexander was engaged in engineering problems of plant extension, radio facilities and special services. With the A.T.&T. Co. he has been engaged in work on data transmission. Member I.R.E.

JACOB F. DEWALD, B.S., 1943, Haverford College; Ph.D., 1949, California Institute of Technology; faculty, Duke University, 1948–49; faculty, Yale University, 1949–52; Bell Telephone Laboratories, 1952—. He has been engaged in studies of ionic and electronic motions across a variety of surfaces, recently concentrating on the semiconductor/electrolyte interface. Member American Chemical Society, American Electrochemical Society, American Physical Society, Phi Lambda Upsilon, Sigma Xi.

J. J. FORST, Newark College of Engineering; Western Electric Company, 1942–48, 1950–55; Bell Telephone Laboratories, 1948–49, 1955—. Mr. Forst has been a member of the mechanical research group since returning to Bell Laboratories, engaged in work on silicon and germanium strain gages and photoelastic stress analysis.

FRANZ T. GEYLING, B.S., 1950, M.S., 1951, and Ph.D., 1954, Stanford University; Bell Telephone Laboratories, 1954—. He was first engaged in research in photoelastic stress analysis and shell theory. More recently his work has been in ballistics of rockets and satellites. Member American Rocket Society, American Society of Mechanical Engineers, International Association of Bridge and Structural Engineering, Society of Experimental Stress Analysis, Phi Beta Kappa, Sigma Xi, Tau Beta Pi.

RICHARD A. GIBBY, B.S., 1949, and M.S., 1950, University of Utah; Ph.D., 1955, Northwestern University; Bell Telephone Laboratories,

1955—. He was first engaged in studies of the acoustical properties of telephone apparatus, and is presently engaged in experimental and theoretical studies of data transmission. Member I.R.E., Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

ROBERT M. GRYB, B.S. in E.E., 1946, University of Illinois; New York Telephone Co., 1946–51; Bell Telephone Laboratories, 1951—. Since transferring from the New York Telephone Co., where he was a traffic engineer, Mr. Gryb has been concerned with traffic measuring devices, No. 5 crossbar switching equipment, and switching and signaling techniques for naval communications. He is at present in charge of a group concerned with data transmission systems. Member A.I.E.E., I.R.E.

FRANK HARARY, B.A., 1941, and M.A., 1945, Brooklyn College; Ph.D., 1948, University of California; mathematics faculty, University of Michigan, 1948—; Bell Telephone Laboratories, 1957–58; Institute for Advanced Study, 1957–59; research staff, Princeton University, 1958–59. While on leave from the University of Michigan, Mr. Harary's work at Bell Laboratories was mathematical research with application to social psychology and sociometry. Member American Mathematical Society, Association for Symbolic Logic, Linguistic Society of America, Mathematical Association of America.

M. E. HINES, B.S., 1940, and M.S., 1946, California Institute of Technology; Pacific Telephone and Telegraph Co., 1936–39; Bell Telephone Laboratories, 1946—. His early work with Bell Laboratories was in development of electron tubes for high frequency and microwave systems. He has been engaged more recently in development of pulse code modulation transmission systems. Mr. Hines' present work is in solid state device research. Member I.R.E., Tau Beta Pi.

DAVID W. NAST, B.E.E., 1953, Cornell University; M.E.E., 1956, Newark College of Engineering; post-graduate studies, Columbia University; Bell Telephone Laboratories, 1953—. His work in systems engineering has included development of a transistorized repeater and transmission studies relating to electronic switching. He is at present in charge of a group concerned with pulse systems transmission engineering. Member I.R.E., Eta Kappa Nu, Phi Kappa Phi, Tau Beta Pi.

IAN C. ROSS, A.B., 1943, and Ph.D., 1957, Columbia University;

Institute for Social Research, University of Michigan, 1949–58; Bell Telephone Laboratories, 1958—. Mr. Ross has been concerned with analysis of the Bell System management attitude survey, and, more recently, with studies of recruiting of laboratory personnel. Member American Association for Public Opinion Research, American Sociological Association, American Statistical Association, Society for the Psychological Study of Social Issues.

W. VAN ROOSBROECK, A.B., 1934, and A.M., 1937, Columbia University; Bell Telephone Laboratories, 1937—. He was first concerned with development of pyrolytic carbon resistors and, during the last years of World War II, with analysis of infra-red bolometers. He later was engaged in studies of characteristics of the copper-oxide rectifier. Since 1948 he has specialized in research in solid state physics, with emphasis on theoretical aspects of semiconductor physics. Fellow American Physical Society; member American Association for the Advancement of Science, American Mathematical Society, Phi Beta Kappa.

BOGUMIL M. WOJCIECHOWSKI, E.E.(M.S.), 1936, Politechnic Institute of Warsaw; Research Staff, Physical Department, Politechnic Institute, 1936–38; Technical Advisor, Polish Stratospheric Board, 1937–39; National Institute of Telecommunication (Warsaw), 1937–39; Research Bureau, Industrièlle des Téléphones (Paris), 1939–40; graduate studies, Sorbonne, 1939–40; Western Electric Co., 1942–55; Bell Telephone Laboratories, 1955–60. At the Western Electric Co. Mr. Wojciechowski was engaged in development of electrical methods and apparatus for precise industrial measurements. After transferring to Bell Laboratories he was engaged in development of special testing equipment for transmission measurements. Senior member I.R.E.; member A.I.E.E. Mr. Wojciechowski died on April 18, 1960.